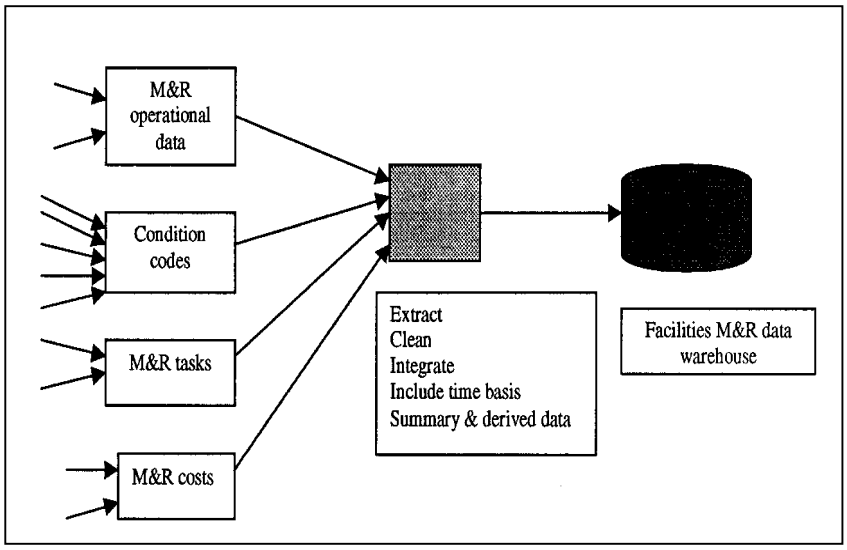
US Army Corps
of Engineers
Engineer Research and
Development Center

CERL Technical Report 99/94
November 1999

Data Warehouse Architecture for Army Installations

Prameela V. Reddy and Charles G. Schroeder



U.S. Army installations are enterprises performing tasks such as management of funds, budgeting, estimating, managing facilities, maintaining facilities, providing training, complying with environmental and safety laws and regulations. Installations use many database management and operational systems to conduct these tasks.

A data warehouse is a single store of information to answer complex queries from management using cross-functional data to perform advanced data analysis methods and to compare with historical data. In the data warehousing approach, the cleansed and transformed data from several operational systems is stored in a single integrated repository of information. This approach provides easy access to needed data, improves system response time, and enhances data integrity. If designed and developed properly, an Army installation data warehouse has the potential to improve efficiencies and produce a positive return-on-investment.

Foreword

This study was conducted for the U.S. Army Corps of Engineers, Directorate of Military Programs (CEMP) under Project 40162784AT41, "Military Facilities Engineering Technology"; Work Unit PL-AH9, "Data Mining for Executive Decision Support." The technical monitor was Leo E. Oswalt, CEMP-IB.

The work was performed by the Business Processes Branch (CN-B) of the Installations Division (CN), Construction Engineering Research Laboratory (CERL). Dr. Moonja Kim is Chief, CN-B and Dr. John Bandy is Chief, CN. The technical editor was Linda L. Wheatley, Information Technology Laboratory.

The Director of CERL is Dr. Michael J. O'Connor.

Contents

List of Figures and Tables

Figures

Tables

1 Introduction

Background

Army installations are enterprises performing tasks such as management of funds, budgeting, estimating, managing facilities, maintaining facilities, providing training, complying with environmental and safety laws and regulations. Installations use many database management and operational systems to conduct these tasks. The current approach used by the Army to support installation management is to develop program interfaces between one application and another to integrate data between applications. The problems associated with this approach of application interfaces are data accuracy, productivity, and the high cost of integration and maintenance.

Traditional operational systems are organized around applications with a focus on functions of limited scope, but strategic decisionmaking usually requires access to and analysis of integrated, historical information. Accurate decision support information, therefore, is not always available to the decisionmaker in the traditional approach. The required raw data comes from many sources and exists in a variety of forms. This data must be cleansed and reconciled to support end users' decisions.

A data warehouse for Army installations is a single store of information to answer complex queries from management using cross-functional data to perform advanced data analysis and to compare historical data. In the data warehousing approach, the cleansed and transformed data from several operational systems are stored in a single integrated repository of information. A data warehouse is a subject-oriented, integrated, nonvolatile, and time variant collection of data in support of management's decisions (Inmon 1996). This approach provides easy access to needed data, improves system response time, and enhances data integrity.

Businesses are using data more effectively with a data warehouse technology to increase profits, reduce costs, and keep their competitive edge. Government organizations are consolidating disparate databases running on incompatible computer systems and forming centralized data repositories that enable quick information retrieval. Army installations can use the benefits of this technology

and improve their efficiencies in managing installations. Identification of trends, clusters in the data, and making forecasts requires analysis of historical information that is not available with existing operational systems. The technology has matured enough so that several commercial tools are available to automate parts of data warehouse development. Data warehousing has come to be seen as a process rather than a product. Deploying a data warehouse or data mart involves not just initial design but also operational processes to populate and maintain it, and to accommodate new data sources.

The data warehousing approach enables data from different sources to be joined and allows new and innovative analysis. Data warehousing helps in managing and maintaining accurate information as a central source of data for cross-functional and historical analysis. It simplifies information maintenance tasks with Army standard data in the integrated database as a data warehouse with clear definitions, data ownership information, source systems, security and control information, business rules, and other metadata.

Objective

The objective of this research was to determine viability of the data warehousing approach for data management of Army installation base operations (BASOPS) functions in support of installation commanders.

Approach

Researchers examined Army BASOPS functional requirements for operations, system integration, and executive decision support. A literature search was performed for data warehousing and data mining. Industry cases studies on data warehousing were reviewed along with demonstrations of data warehousing, data mining, and data visualization tools. CERL worked with the U.S. Army Center for Public Works (now the Installation Support Division) and Directorate of Public Works (DPW) at Fort Eustis, VA, to develop a conceptual data warehouse for the DPW functional domain.

Research results include the problems and difficulties with the current data integration approach and the status of data warehousing technology (Chapter 2). The scope, structure, types of data to include, and overall architecture of data warehousing is explained in Chapter 3. Data warehouse design and development methods are discussed in Chapter 4, including some popular commercial tools to help simplify the development process. Chapter 5 describes decision

support technologies to help end users with a data warehouse in place. On-line analytical processing (OLAP) and data mining technologies are discussed in detail in relation to Army installation decision support. Some of the popular commercial tools for decision support are listed. Data warehousing technology and decision support technologies are discussed in detail in Chapter 6 with a focus on Army installations.

Mode of Technology Transfer

The work reported here will be shared with the Corps of Engineers Installation Support Division (ISD), the Assistant Chief of Staff for Installation Management (ACS(IM)), and U.S. Army Strategic and Advanced Computing Center (SACC). ISD and CERL will work together to plan and implement a data warehouse for managerial decision support at the installations. Data consistency with the Headquarters, Department of the Army (HQDA) data warehouse will be maintained by sharing the installation data warehouse approach with SACC. These methods and techniques are applicable to DPWs at other military installations.

2 Comparing the Traditional Approach to the Data Warehousing Approach

Today's rapidly changing business environment demands increasing amounts of timely information to support decisionmaking needs. Many organizations are becoming increasingly customer focused and recognize that their large databases of customer-related data can be analyzed to extract information for business advantage.

Traditional Approach

Because of the way data applications have been and continue to be built, they not only contain data divorced from a business context but also data seldom consistent across the breadth of the organization. The application-oriented approach to delivering data to end users is restricted in scope and highly localized. Users are primarily interested in what can be done with the data on hand rather than what other, better data are available.

Most data applications are developed to increase speed and accuracy of running a business. They are called operational systems. They focus on separate areas of business function (e.g., budget, personnel, facility maintenance) with well-defined needs. This focus enables more rapid deployment of functionality. In reality, however, the sets of data in operational systems are not truly independent. They are transaction-processing applications that create and use several databases. Some applications overlap or use information from other databases or operational systems. The required raw data comes from many sources, both internal and external to the business organization, and exists in a variety of forms, from traditional structured data to such unstructured data types as documents or multimedia. The data must be cleansed and reconciled to ensure its quality and integrity. Decisions are based on the combined set of data with interlinked and overlapping data domains.

This interlinked environment creates a spider web of extract programs. Figure 1 shows a spider web of application interfaces at Army installations. The problems and difficulties associated with this spider-web architecture are credibility

of data, productivity, and inability to transform data into information. Lack of credibility results from differences in the time basis of data, algorithms, levels of extraction, external data, and fundamental definitions of the data (Inmon 1996). The productivity of the organization is adversely affected by extract programs that are customized and cross many technological barriers. The customized programs will take a very long time to accomplish tasks and a large amount of resources. The systems found in the naturally evolving architecture are simply inadequate for the task of supporting information needs. They lack the integration and historical data needed for decision support. These systems are also inadequate because of the differences in the time horizon of applications.

In some cases, direct access to legacy databases is provided through a common front-end to multiple systems. Some existing consolidated databases are limited by older mainframe technologies, in which pre-formatted queries are established with user-defined variables. Some organizations have moved data to client/server platforms because of the difficulty of accessing the data in existing non-relational legacy systems. In other cases, although data integration of disparate systems is taking place, the full power of using data for analytical purposes has not been explored.

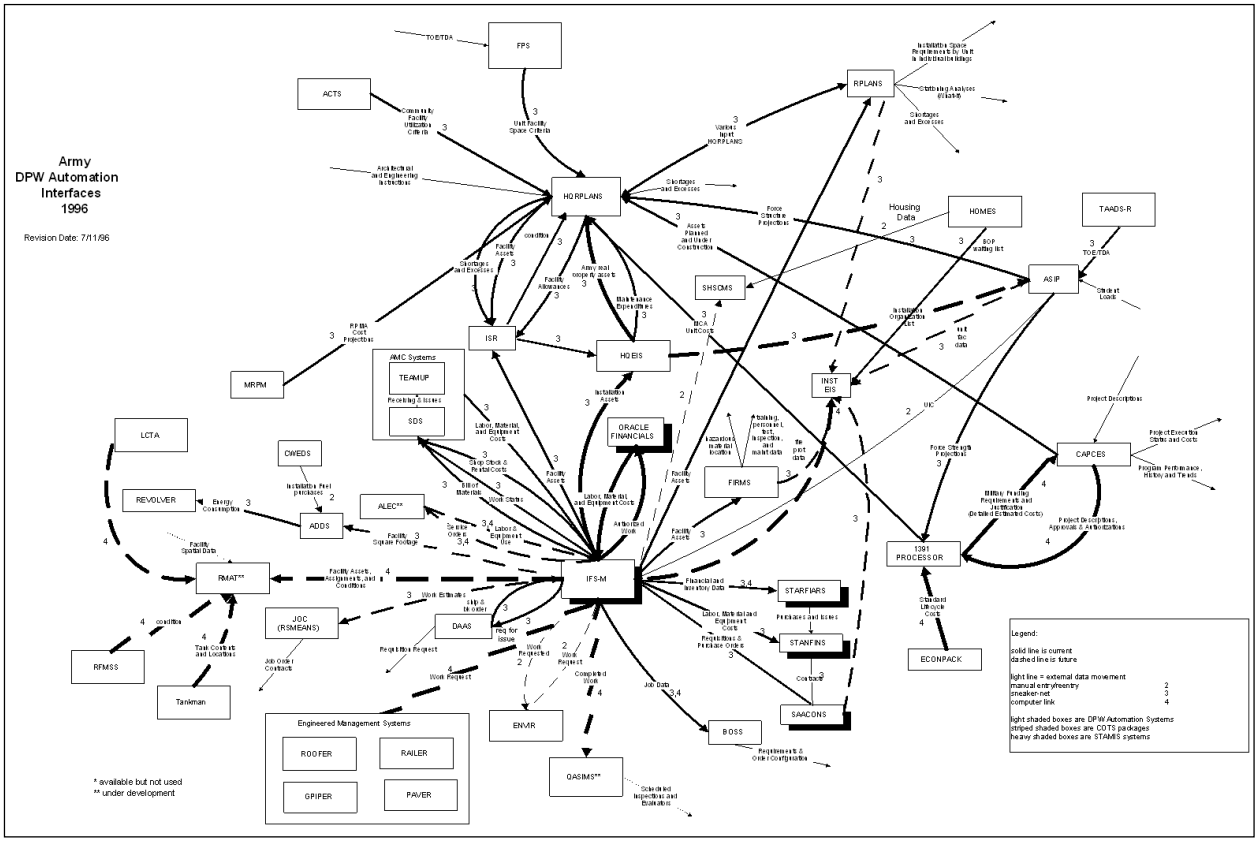


Figure 1. Army DPW automation interfaces (1996).

Strategic business decisionmaking usually requires access to and analysis of integrated, historical information in a timely manner. Although business organizations have been generating great quantities of transaction data for years, potential uses for these data have not been fully exploited because of the difficulty in consolidating the data into useful information. There is a need to look across the organization to support integrated business processes as the organization pursues business process reengineering. Developing an efficient and effective database architecture to integrate operational data will help organizations gain strategic advantage and realize business opportunities. Many organizations look to data warehousing to meet this challenge.

Army installations use many database management systems and operational systems to conduct business. Development of applications for Army installations follows the same traditional approach with limited scope and well-defined requirements. Applications are developed focused on separate functional areas. They use and create several databases with limited scope. The Army currently supports installation management by developing program interfaces between one application and another to integrate information between the applications.

Besides the problems and difficulties associated with a spider web of extract programs, Army installations face other problems. The Army is moving towards the use of commercial, off-the-shelf (COTS) software, which presents a new problem in how to manage Army corporate data independent of vendor modifications to their database schemas as a result of product upgrades. The Army is also tending towards privatization of many services. This trend poses a new information technology challenge to determine how to share information practically with external customers, service providers, and privatization partners when they use information management systems that differ from the Army's systems. Data extraction from these third party systems can be much simpler with data-warehouse architecture than with traditional spider-web architecture.

Data Warehousing

Computers are used to process transaction data and to provide information to support decisionmaking. Data warehousing recognizes the merits of placing specially prepared data on separate computer platforms for decision support purposes. Decision support systems (DSS), executive information systems (EIS), and many other applications benefit from having these separate platforms. This approach provides easy access to needed data, improves system response time, and enhances data integrity. Enhanced data access tools make it easier for end

users to access, analyze, and display information without having to know, for example, how to write Structured Query Language (SQL) queries.

Data warehousing enables data to be joined from different sources and allows new and innovative analysis. A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use in a business context (Devlin 1997). Data warehouses provide decisionmakers with information that accurately and effectively reflects the entire business.

Data warehousing is one of the most rapidly growing areas in management information systems, separating data for DSS and EIS applications from operational data, and storing it in a custom-designed database. Instead of an application with which users interface directly, though, a data warehouse is an integral part of an organization's underlying technical infrastructure. This approach results in improved performance, better data quality, and an ability to consolidate and summarize data from heterogeneous legacy systems.

In a typical organization, operational data is scattered throughout a variety of database management systems using widely different formats and hardware platforms. Accessing this data and making it available for DSS and EIS applications is often difficult and time consuming. The realization that there is not enough time or money to replace legacy systems, together with the ever-increasing demand for more data that is more reliable, has caused data warehousing to grow in popularity in the information management area.

Several highly successful data warehouses are developed and deployed for businesses of all types and sizes. Data warehouse tools have evolved to the point where it is often economically feasible for even small firms to construct and deploy them. Advances in data modeling, databases, data mapping tools, and application development methods make a data warehouse feasible as a primary data source for executives, managers, analysts, and knowledge workers.

A 1996 study conducted by International Data Corporation reported that 62 organizations that had implemented data warehouses showed an average return on investment of 401 percent over 3 years, and the average payback for the warehouse application was 2.3 years. Half of the organizations reported returns greater than 160 percent and one quarter showed returns greater than 600 percent Yet these numbers only scratch the surface of a data warehouse's true value. The true and most far reaching benefit of a data warehouse lies in the solid decisionmaking it enables.

Summary of Design

The design of a data warehouse should be based on an analysis of user requirements. It does not exist in isolation but is part of a larger client/server environment. The extent to which the data warehouse is used suggests how well it meets organizational needs. The following categories of data are typically maintained in a data warehouse: current detail, historical detail, lightly summarized data, and heavily summarized data (Inmon 1996). Data in the warehouse cannot be independently updated by users, but rather is refreshed on a periodic basis by data extracted from various data sources. A facility with the means for removing and archiving aging historical detail data must also be provided. This capacity prevents the unbounded, uncontrolled growth of the warehouse, which ultimately could overwhelm a system.

An active repository is a critical component of the data warehouse architecture. The repository houses metadata that indicate where data comes from; how it should be translated or transformed in order to move it to the data warehouse; who accesses the data and how often; what business processes it drives; and which critical success factors it supports. The repository supports the building and maintenance of the data warehouse. It is also essential in supporting end users in accessing and analyzing data.

Data warehouse development requires top management support and is generally managed as an iterative process. The typical data warehouse architecture today is a two-tier architecture supporting an enterprise-wide community of EIS and DSS clients. However, some organizations are evolving a three-tier architecture, with an additional server layer inserted between the data warehouse and the user community. The purpose of the new server layer is to facilitate the creation of user-community-specific data marts that focus on end-user requirements for data. A subset of data is extracted or summarized from the data warehouse and is optimized for each type of user. In essence, the data warehouse acts as a "wholesale" source of data, and that data is "retailed" to the data marts based on local need.

Data Warehouses in the Private Sector

The adoption of data warehouses has helped many companies respond to an ever shifting competitive environment. Simply put, a data warehouse is just another database. What sets it apart is that the information it contains is not used for operational purposes, but rather for analytical tasks – everything from brainstorming to identifying new methods, forecasting future capacity, future supply

and demand, and value management. Most large companies have installed data warehouses, or are in the process of doing so. Some use it to identify purchasing patterns of customers and others use it to rationalize inventory and supply. Still others use it to forecast demand for their products.

In 1995, Wal-Mart deployed its data warehouse to support its decisionmaking. Its retail stores around the country pour daily transaction data into its data warehouse, from which Wal-Mart can analyze what is selling, where it is selling, and when it is selling. Its data warehouse helps managers in various functional areas determine optimal pricing and inventory levels, as well as the most effective way of promoting each store's products.

VF Corporation, a textile company in North Carolina, created the market response system that analyzes data from retail stores, which is stored in its data warehouse. This information is used to maintain inventory down to the style, color, and size to share with designers, fabric buyers, manufacturers, and retailers. This market response system reduced the conventional 100 to 125 day product development cycle to about 35 days.

Advocate Health Care in Oak Brook, IL, a leading provider of health services in the Midwest, is using a data warehouse to assess patient care quality and to respond to patient needs more effectively. It also uses its data warehouse to predict market needs for organizational planning, including customer and staffing needs.

Piedmont Hospital in Atlanta, GA, a 500-bed facility, built a client/server data warehousing application with financial and patient information. This information is used to track and control costs, and to determine the cost of treating patients and the effectiveness of those treatments.

Many Wall Street investment banks and securities firms realized that islands of information systems made it nearly impossible for them to assess risk. They began building data warehouses as a way of collecting operational data and rearranging the data in a centralized relational format. With data warehousing, brokers and dealers now analyze risk and distribute that information. In addition, investment firms such as Sumitomo Bank and Prudential Insurance Co. have finished building data warehouses to analyze market movements, assess company-wide risk, manage portfolios, track customer tendencies, and settle trades. With their data warehouses, these firms now can concentrate on analyzing the data and responding more swiftly to market changes and customer needs.

British Air uses its data warehouse to estimate its route profitability and to deploy its fleet efficiently. Another example from the airline industry can be found in USAir's frequent flyer program. USAir originally tried to create a data-mining system that accessed the company's on-line transaction processing (OLTP) environment directly. As soon as they started data mining, however, they slowed down the transaction processing rate, and realized that a data warehouse was needed to meet their information needs.

Data Warehousing in the Public Sector

Government organizations, like those in the private sector, are either building data warehouses, considering building them, or involved in a transition from older technologies to client/server technology. As in the private sector, government agencies are looking for ways to make dramatic improvements in work practices. These reengineering efforts often require supporting integrated business processes that cut across existing organizational lines. Data warehousing is a mechanism to consolidate data from departmentally focused systems into a unified database that supports a cross-organizational viewpoint.

Data warehousing, with integrated and accessible data, allows for fast responses to external requests that may otherwise require extensive manual efforts to create specialized reports and reconcile disparate and often inconsistent data. Data that were previously inaccessible now can be accessed and compared across subject areas that were previously difficult to execute because of standalone transaction systems. A single, consistent source of high quality historical data can support trend analysis and forecasting, identification of key events, and data mining to discover patterns of behavior. Data warehousing provides a mechanism to improve the quality of data by defining common data structures and formats, and enforcing consistent data domain values through data transformation.

The SACC is developing a data warehouse that will support HQDA with consolidated data on Army units, personnel, logistics, facilities, readiness, and budget. This data warehouse will reduce costs associated with duplicative data acquisition, reconciliation, and integration efforts; improve the quality and consistency of data; promote data sharing; and make data easily accessible to users. The components of this data warehouse include integrated data, metadata repository, data access tools, data transport and cleaning tools, and data models of all the data contained in the warehouse.

The U.S. Environmental Protection Agency (EPA), during its reengineering efforts, decided to provide environmental information to the public with a focus on

geographic regions. Instead of several databases focusing on air, land, waste, water, and toxic substances, they needed an integrated database to study all elements affecting the environment in an area. A centralized data warehouse was built using information from numerous databases and linked to the World Wide Web so that information could be made available to the general public. Their website "Envirofacts" reports on everything from air pollution levels to hazardous waste site assessments.

Other government organizations include the Naval Surface Warfare Center, which built a data warehouse that cleans and consolidates data from many different sources to address management decisionmaking in the areas of human resources, finance, customer service, procurement, and organizational structure. The Department of Housing and Urban Development's web warehouse integrates community planning and housing project information from various sources and arranges it on maps. The Transportation Department uses its web warehouse to share information with other government agencies. The Bureau of Labor Statistics makes wage, price, and employment data publicly available.

Within the commercial sector, the major driver for data warehousing relates in some way to maximizing profits in a competitive environment. Profitability is generally not a factor in the government sector, but most types of organizations are concerned with issues of mission obligations, budget, personnel, performance, tasking, and evaluation. All of these issues are at least partly data-driven and data-measured. Privacy and security considerations are somewhat different in the government sector and should be considered when building a data warehouse.

A research-oriented organization (other than a vendor) needs to conduct a comprehensive survey of state-of-the-art data warehousing in the government to address areas such as practices, standards, tools, user acceptance, and documented experiences.

Accuracy and Integrity of Data

Within the government sector, initiatives have been implemented to improve the quality of products and services. This emphasis on quality has called attention to disparities in data and the difficulties in obtaining accurate and timely answers to questions. Data warehousing provides a mechanism to improve the quality of data by defining common data structures and formats and enforcing consistent data domain values through data transformation. Data consolidation

efforts also focus attention on data entry practices in source systems, leading to improved editing of data at the source.

The overall objective of the data warehouse is to make high quality, reliable data widely available and easily accessible to all levels of users. It provides an integrated single source of information that is time-synchronized. As a single source for enterprise data, it provides consistent data. The data warehousing process involves collecting, consolidating, organizing, transforming, and storing data in a database management system environment so it is available to users. This process cleans legacy data to make it consistent with enterprise-wide standard data elements. Another benefit of this process is standard data element visibility because the legacy data elements are mapped to the standards.

Most of the Army's applications were designed to support specific functions such as finance, supply, or procurement — usually at the national level. These systems are not integrated with one another nor are they designed to give the local commander essential management information (ICIM 1995). Despite an array of systems, many do not follow standard data elements and lack interoperability. One of the major functions of the Department of Defense (DOD) Corporate Information Management (CIM) program is enterprise information integration. The DOD enterprise model is a representation of the DOD's activities and data. It is the basis for defining, coordinating, and integrating DOD missions and functions. Data modeling is a key aspect of the DOD data administration program and drives data element standardization.

Standard data elements are the key for accuracy of the data. If one department has extracted its data for analysis in January and another department doing analysis has extracted its data in March, the data will not be same and management is faced with making decisions with inconsistent data. Creation of data at different points in time is only one problem. There may also be algorithmic differences between the analysis methods of the two systems. A facilities condition assessment method in one department may differ from the condition assessment method used by another department. Integrating data directly from legacy systems will not be accurate. Differences may exist in data definitions, naming conventions, formats, and the time basis. A data warehouse with accurate metadata will provide more accurate information. Both the users and source data owners must have a high level of confidence in the validity of the data. A detailed and accurate metadata repository in a data warehouse provides accurate and extensive knowledge about the data in the warehouse.

Historically, most business computing has been directed toward operational systems. During the development of operational systems, users perceive a set of

functions as related and applications are developed in some integrated manner to support that process. However, the scope of data included in these operational systems is limited. Figure 2 shows example data scopes of operational applications in a traditional development environment.

While this view is adequate for the users, in reality, these sets of data are not truly independent. In some instances, data flows from one domain to the other, while in other cases, data is shared between domains. Figure 3 expresses how data are shared between operational applications. Common data are stored and accessed in one or more shared databases. This approach enables a high level of consistency between the data in these applications. As demonstrated in Figure 3, application-oriented development is optimized to deliver systems to users with clearly defined but narrow goals. This can, and often does, lead to a situation where application designers build their own data definitions, based on their local application needs. Problems arise when data are inconsistent between the known applications and other applications from other departments.

The inconsistencies that exist and will continue to exist in operational data mandate that a single source of informational data be defined. This single source would prevent operational inconsistencies from being reflected in the informational world. In reality the development of operational applications is function-driven rather than data-driven and is likely to remain that way for the foreseeable future. As a consequence, data in the operational environment will continue to lack the required level of consistency to enable true cross-enterprise data use.

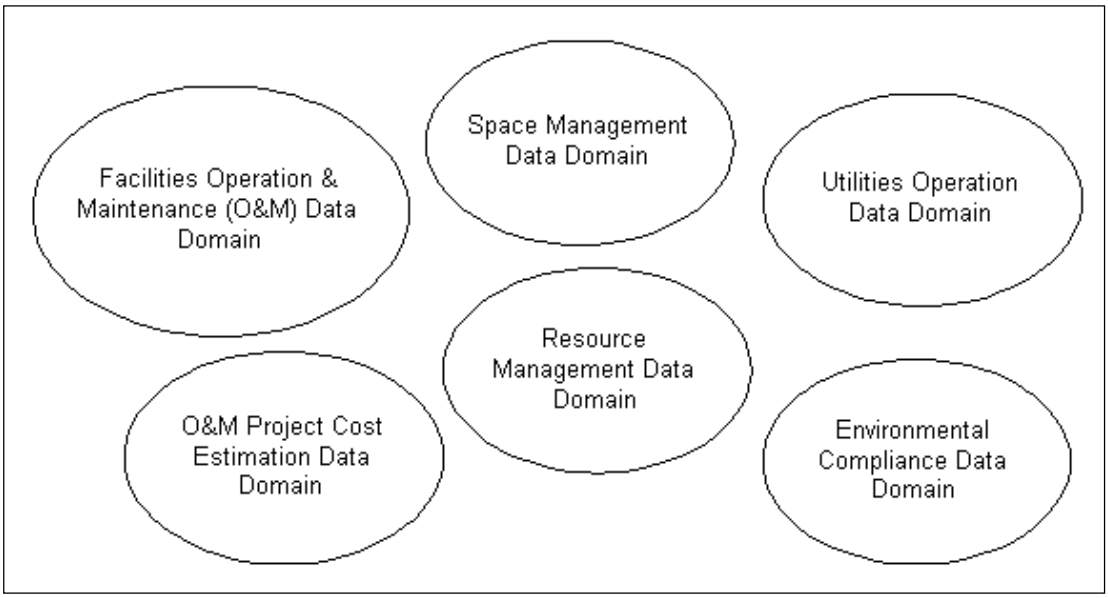


Figure 2. User view – independent data domains.

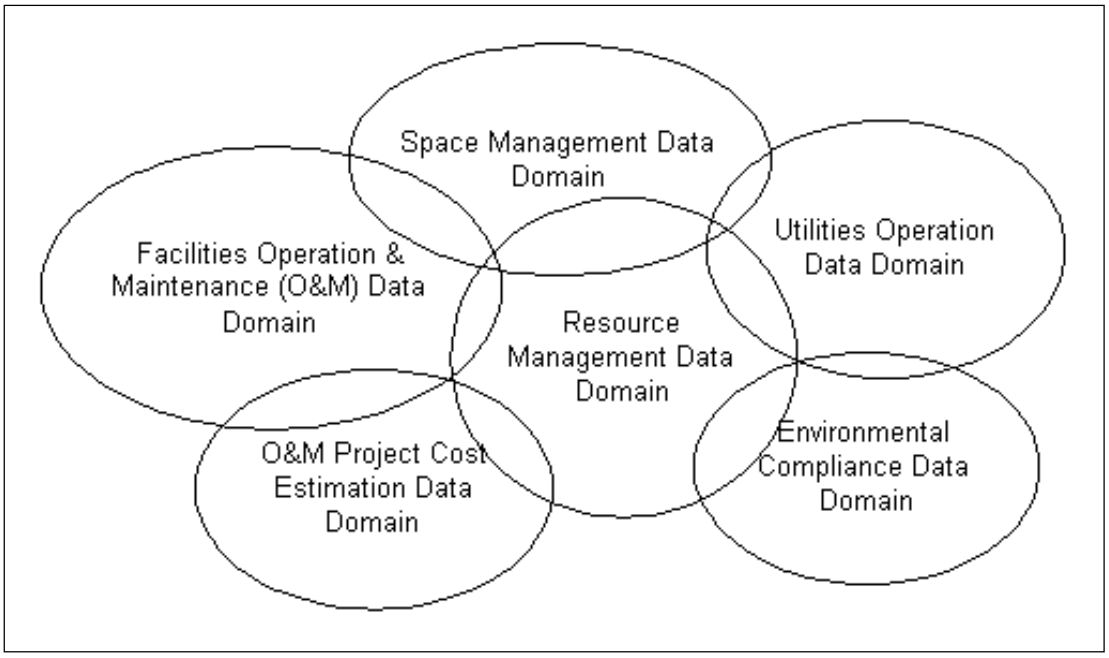


Figure 3. Reality – interlinked and overlapping data domains.

Since informational systems developers were familiar with application-driven development for operational systems, it was only natural that they would use the same approach when the first demands for informational applications emerged. Vertical fragmentation of informational systems is typical in many organizations. This fragmentation is clearly business-driven because the majority of the decision support needs in a department are related to data originating within that department. Users tend to express their needs for informational applications in terms of automating the delivery of familiar reports. As end users' familiarity with the data available to them increases, so does the sophistication of their requirements. As the analysis needs of the business change, the copy programs become increasingly complex, and maintenance becomes a major problem.

Another reason for the increase in the complexity of copy programs occurs when users begin to recognize some inconsistencies in the data they receive. Solving the problems at the source is often too expensive or too slow, and they may not own the data to fix it in the operational system. In these cases, the copy programs may include corrections, further increasing their complexity and their maintenance costs. Given these problems, copy programs have a tendency over time to become a "spider web" of code. Even within individual departments, users sometimes need data from more than one operational application and then must combine the results. Figure 1 is an example for such data integration requirements in DPW. Master planners at DPW use the IFS-M system, 1391

processor, Econpack, FPS, ACTS, and several other systems to analyze facilities requirements and to plan for future growth.

System Maintenance Costs

Data and information resources presently consist of scores of disparate legacy systems that are located on a wide variety of applications, operating systems, and platforms. The nature of these data systems is that they were designed as standalones with little or no consideration for interaction or compatibility with other information systems. Trying to draw consistent information from them on a common basis is almost impossible.

Sometimes updated and outdated data are combined, which produces erroneous results. A change in any data item in any application can affect all of the users. The application developers are not always aware of all the data links. If any link in the chain fails, subsequent links are no longer valid. Any changes in the structure of the base operational data can cause substantial duplication of maintenance effort on the related copy programs. This results in very high maintenance costs if the maintenance costs of all applications in any functional area are combined. Efficient use of these data systems is further limited because of the complexity of their various data structures. Often, no information is available for the user to understand the data structures or the data values.

In addition to these data inconsistency problems, the technical difficulties of obtaining data from different hardware and software platforms are also a significant obstacle to overcome. Temporal inconsistency in the data is another problem. Data in different applications sometimes have different time spans and so cannot be directly combined. If users recognize this data inconsistency problem and problems with copy programs, they may duplicate some needed data. This creates another problem. Data is duplicated many times, often in hidden or forgotten ways. These duplicated efforts are very resource intensive in time, personnel, hardware, and software. The proliferation of existing and future legacy systems, coupled with the different ways each staff element accomplishes its data integration and synchronization, greatly diminishes the degree of consistency in data and increases the resource requirements for data acquisition, reconciliation, and integration.

Inmon (1996) wrote about three main problems with the current approach:

- credibility of data
- productivity
- inability to transform data into information.

Credibility of Data

Credibility of data is the most important objective for any organization, since all of the managerial decisions are based on the data available to that organization. Five reasons for the lack of data credibility are:

- no time basis of data
- the algorithmic differential of data
- the levels of extraction
- the problem of external data
- no common source of data to begin with.

According to Inmon, a crisis of credibility is brewing in every organization that allows the formation of a spider's web with its software and data.

Productivity

The second major problem with current architecture is its effect on productivity. The process of going through every piece of data — not just by name but by definition and calculation — is very tedious. Unless data are analyzed and rationalized, the report will end up mixing apples and oranges, creating confusion. Writing a program to extract data from many sources for one application may not be difficult. It becomes complicated, however, if there are lots of programs for several applications, and each is customized. Unless future data requirements are known in advance, and unless those requirements are factored into the report generation program, there is every likelihood that each new report will have to pay the same large overhead.

Data Transformation and Integrity

Data integrity drives business reengineering and is a fundamental issue in data warehousing. The current architecture of systems prevents the transformation of data into reliable information. Finding the information for decision support from existing operational systems is very difficult and almost impossible. With several applications, several databases, and several levels of detail, trying to draw information from them on a common basis is almost impossible. The applications were never constructed with integration in mind. Obtaining any useful information from across the different data systems requires extensive data collection, synchronization, and integration efforts that are often redundant and consume considerable amounts of scarce resources such as money, personnel, and time. A data warehousing environment with high quality, reliable data widely available and easily accessible to all levels of users benefits both the customers and custodians of legacy systems.

DOD and Army recognized this problem and began a data standardization effort to manage data as a strategic resource. DOD's CIM integration architecture contains seven levels from a personal level to a global level. Application, function, mission, and enterprise levels are different levels of integration with security barriers. Each level in the architecture has a distinct set of objectives, requirements, methodologies, techniques, and tools embedded. The CIM integration architecture provides a framework that guides integration of legacy systems. The data warehouse method can be used to implement this strategy using information integration architecture.

One of the major tasks of developing a data warehouse is data cleaning, which involves determining data elements and attributes, standardizing, verifying, matching, and documenting. A data warehouse significantly reduces the duplication of data collection and preparation efforts and also reduces the data replication in the source systems by providing a single acquisition source for consistent, authoritative, and easily accessible data.

3 Data Warehouse Architecture Overview

A higher level definition of what constitutes a data warehouse is appropriate before going further on the architecture of a data warehouse. There are a few definitions from data warehousing experts: A data warehouse is a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management's decisions (Inmon 1996). A data warehouse is a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use in a business context (Devlin 1997). A data warehouse is designed for strategic decision support, and is largely built up from the databases that make up the operational database (Adriaans and Zantinge 1996).

For purposes of this report, a working definition of the data warehouse is that it is a collection of integrated, subject-oriented databases designed to support decisionmaking activities, where each unit of data is relevant to some moment in time. The classic system development process starts with the identification of system requirements. To build systems, the requirements must first be understood. Then, design and development begin. A data warehouse's change in architecture starts with data. Requirements are usually the last thing to be discovered. Once data from several sources are cleaned, integrated, and a data warehouse is implemented, programs are then written against that data.

At the core of the data warehouse environment is the realization that there are fundamentally two kinds of data – operational and derived. Operational data is detailed to run the day-to-day operations of the business. Derived data has been summarized or otherwise calculated to meet the needs of management. Operational data is primarily current value data that can be updated. Derived data is often historical data that cannot be updated.

The data warehouse holds time-based operational data and some derived data. As data pass from the operational environment to the data warehouse environment, they are integrated. When the data need to be brought together from more than one source application, it is natural that this integration be done somewhere independent of the source applications. The data warehouse very effectively combines data from multiple source applications such as facilities inventory, maintenance and repair (M&R), cost estimation, utilities operation,

environmental compliance, and master planning and programming. Many large data warehouse architectures allow for the source applications to be integrated into the data warehouse incrementally.

The primary reason for combining data from multiple source applications is the ability to cross-reference data from these applications. Nearly all data in a typical data warehouse is built around the time dimension. Time is the primary filtering criterion for a large percentage of all activity with the data warehouse. For example, one may compare the condition of one type of facility for this year with the condition of that type of facility for prior years. The time dimension in the data warehouse also serves as a fundamental cross-referencing attribute. For example, an analyst might want to compare current environmental compliance data and violation notices with those of previous years. Management may attempt to assess the impact of business process reengineering activities by comparing current results with previous years' results.

Another key attribute of the data in a data warehouse system is that the data are brought to the warehouse after they become mostly nonvolatile. This means that, after the data are in the data warehouse, no modifications are to be made to this information. For example, the condition of buildings does not change, the building inventory snapshot does not change, and the M&R cost details do not change for any particular time dimension. In an operational system, the M&R data entities and attributes go through many changes. For example, the status of a work order may change many times before the work is completed. Another example is a product moving through the assembly line that has many processes applied to it. Generally speaking, data from an operational system are triggered to go to the data warehouse when most of the activity on these business entity data has been completed. This may mean completion of a work order or M&R cost for that work order. Inventory may change with every transaction and it is impossible to carry all of these changes to the data warehouse. A snapshot of inventory for a specific time period, determined by management and carried to the data warehouse, is sufficient for all analysis.

The cost of maintaining the data once it is loaded in the data warehouse is minimal. Most of the significant costs are incurred in data transfer and data scrubbing. For this reason, storing data for more than 5 years is very common for data warehousing systems. The separation of operational data from analysis data is the most fundamental data warehousing concept.

The Structure of the Data Warehouse

The data warehouse has three different levels of detail: a current level of detail, a level of lightly summarized data, and a level of highly summarized data. Data flows into the data warehouse from the operational environment. Significant transformation of data usually occurs during the passage from the operational level to the data warehouse level.

As stated earlier, storing data for more than 5 years is very common for data warehousing systems. As data ages, it passes from current detail to lightly summarized data, then from lightly summarized data to highly summarized data. It may be transformed from daily transaction data to monthly data, and then to yearly data.

The data model outlines the logical and physical structure of the data warehouse, which is oriented to the major subject areas of the business that have been defined in the data model. This data modeling process needs to structure the data independently of the data models that may exist in any of the operational systems. Some elements and attributes that are essential to the operational system may not be necessary for the data warehouse. A data warehouse project in most cases cannot include data from all possible applications right from the start. The project is designed and the data populated one step at a time.

The data are logically transformed when brought to the data warehouse from the operational systems. The architecture of the data warehouse and the data warehouse model greatly affect the success of the project. Therefore, the data modeling effort in the early phases of the data warehousing project can yield significant benefits in the form of an efficient data warehouse that can expand to accommodate all of the business data from multiple operational applications. The operational systems, however, are likely to have large amounts of overlapping business reference data that needs to be consolidated in the data warehouse system, leaving only the data relevant for the analysis processes.

The data warehouse logical model aligns with the business structure rather than the data model of any particular application. The data warehouse would most likely build attributes of a business entity by collecting data from multiple source applications because the structure of the data in any single source application is likely to be inadequate for the data warehouse. Physical transformation of data homogenizes and purifies the data. For example, the terms and names used in the operational systems are transformed into uniform standard business terms and definitions by the data warehouse transformation processes.

Issues associated with default and missing values are also managed while moving the data to the data warehousing system. It is important to devise a mechanism for users of the data warehouse to be aware of these default values and missing data.

Many queries and reports run in most data warehouse systems are simple aggregations or summarizations based on predefined parameters. Business view summarization of data is a key attribute of today's data warehouses. The single most important reason for building summary views is the significant performance gains they facilitate. Summary views are able to perform the most time consuming data analysis before it is needed, and are often generated not only by summarizing the detail data but also by applying business rules to it. In addition to applying the business rules while generating summary views, the data warehousing system may perform complex database operations such as multitable joins. These summary views need not only to be designed and built, but to be maintained as new data come into the data warehouse.

Some of the activity supported by a data warehouse is predefined and not much different from traditional analysis activity. Other processes such as multidimensional analysis and information visualization are not available with traditional analysis tools and methods. The standard reports required by many users and predefined summary views account for the majority of activity in a data warehouse. It is desirable to periodically and automatically produce these standard reports that are required by many different users. When these users need a particular report, they just view it because it has already been run by the data warehouse system. They do not need to run it themselves. This facility is particularly useful for reports that take a long time to run.

Even though the standard reports and queries are adequate to answer many questions, answers to "why" and "how" questions are not available in them. Data mining* in the detail data can provide some of these answers. A data mining user starts with summary data and searches or "drills down" into the detail data looking for arguments to prove or disprove a hypothesis.† The tools for data mining are evolving rapidly to satisfy the need to understand the behavior of business units such as customers and products.

---

* Data mining is more fully described on page 52.

† To "drill down" is to move down the levels in a hierarchy, while to "roll up" is to move up the hierarchical levels.

A data warehouse may feed data to other data warehouses or smaller data warehouses called data marts. (For further discussion of data marts, see page 48.) As the data warehouse becomes a reliable source of data, many applications find that a single interface with the warehouse is much easier and more functional than multiple interfaces with the operational applications. Of course, all analysis run at a data warehouse is simpler and cheaper to run than through the traditional methods. This simplicity continues to be a main attraction of data warehousing systems. A flexible enterprise data warehouse strategy can yield significant benefits for a long period.

Data Warehouse Architecture From a Client/Server Perspective

The data warehouse and its supporting hardware and software platforms constitute a large database server that supports an enterprise-wide community of end users. This type of architecture is common in data warehouse applications. It is a two-tier architecture from a client/server perspective. However, some organizations are evolving a three-tier architecture, with an additional server layer inserted between the data warehouse and the user community.

The purpose of the new layer is to facilitate the creation of user-community-specific data marts that focus on end-user requirements for data. In a three-tier architecture, the data warehouse acts as a “wholesale” source of data, and that data is “retailed” to the data marts based on local need. Data on each local server can be stored in the form of a relational database. However, when a three-tier architecture is used, it is common to structure some of the data in the form of a multi-dimensional database.

The main advantages of the two-tier architecture are its simplicity and lower costs. The main advantages of the three-tier architecture are faster response times and the ability to custom design data for each type of user community. There is disagreement in the industry on the issue of two-tier versus three-tier architecture. The major factors that impact this issue include the size of the data warehouse, the number of actual and prospective users, and the types of analyses that are performed.

The Scope of Data Warehousing

The data warehouse provides the well understood business data needed to manage the business. It does not have to cover all the data in the enterprise. It focuses on business data and metadata that are mainly public in scope and covers

both structured and unstructured components of business data and metadata. It is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management decisions.

Given that the data warehousing scope has been defined as management of the business, real-time data lies outside the scope of the warehouse because significant transformation of the data occurs when the data passes from the operational level to the data warehouse level. Personal business data and metadata are also largely excluded from the warehousing scope. Derived data is used to manage the business, making it part of the data warehouse. Other business data that is needed to manage the business falls within the scope of data warehousing. Build-time metadata, used in application development, lies beyond the scope of the data warehouse. Other metadata, however, can be part of it. Reconciled data, a special category of derived data, provides the means to ensure data consistency, which is key to data warehousing. Special measures are usually necessary to ensure quality of external data in a data warehouse.

The data may be highly structured, consisting of many well-defined interrelated fields or records, or unstructured, where the internal structure is variable. Both structured and unstructured data fall within the scope of a data warehouse. Image, audio, and video are examples of highly unstructured data. The importance of less structured types of data is rapidly increasing in all businesses and consequently in informational systems. Any required reconciliation of unstructured data occurs through its associated structured data. A textual account of a traffic accident stored in a text-processing system could be reconciled with a video of the accident scene stored in an image database. Both of these are unstructured data and reconciliation occurs through the claim number, which is structured data.

Unstructured data can and should be included in the data warehouse. However, unstructured data is more voluminous than structured data, and more difficult to manipulate. Therefore, although unstructured data is of considerable business value, structured data is usually implemented first in a data warehouse.

The data in a data warehouse are organized by major subject areas of the business that have been defined in the data model. The collections of data that belong to the same subject area are tied together by a common key that includes both summary data and a level of detail for the data to support management decisions.

The data in a data warehouse are integrated when data pass from operational systems to the data warehouse. The data from different operational systems

may not be consistent in encoding, naming conventions, attributes, measures, etc. Data are entered into the data warehouse in a way that the many inconsistencies at the application level are undone.

Another important characteristic of a data warehouse is that it is nonvolatile. Update of data does not occur in the data warehouse environment. Data updating is done in the operational environment. Data in a data warehouse is a stable snapshot of the business data at a particular moment in time and reflects the status of the business at that moment.

The last characteristic of the data warehouse is that it is time variant. The key structure of the data warehouse always contains some element of time. A 5- to 10-year time horizon of data is normal for the data warehouse. Operational databases contain current data so the current value of the data can be updated. Although the key structure of operational data may or may not contain some element of time, data for a specified time period (determined by management) is carried to the data warehouse.

Types of Data

Many varieties of data are stored in computers today. Devlin (1997) categorizes three main types of data — business data, data as a product, and metadata.

Business Data

Business data is required to run and manage the business. It is created and used through transaction processing systems and decision support systems. Business data can be categorized as both operational and informational data. Operational data is used to run the business day to day. Informational data is used to manage the business in the long term. Operational data is the primary business data within the organization and is the source of all informational data. The value of business data lies in how well it reflects the reality of the business activities. Operational data includes detailed data in real time.

Operational data

Operational data is critical for running the business and is related to short-term actions or decisions. It focuses on transactions such as products, customers, or work orders. *Summary data* is used in managing the business and showing a broad view of how the business is operating. *Real-time data* gives a view of the business at the present time. *Point-in-time data* is a stable snapshot of the

business data at a particular moment in time and reflects the status of the business at that moment. Monthly or yearly data are examples of point-in-time data. Such data can represent views of the past and may be used to predict future events.

This data may be highly structured or unstructured, but management information systems have traditionally focused on well-structured data. The importance of less structured types of data, however, is rapidly increasing in all businesses. Image, audio, and video are examples of highly unstructured data. Any required reconciliation of unstructured data occurs through its associated structured data.

Derived data

Real-time data is processed to produce derived data. Derived data has traditionally been used for decision support. It may be summarized data, new data derived from some combination of existing fields, or a snapshot of detailed data with a time segment attached to it. Reconciled data is generated by a process designed to ensure internal consistency of the resulting data, so reconciled data is seen to be a special category of derived data. Whenever data from multiple sources have to be combined in a data warehouse, data reconciliation is necessary.

Personal data

Personal data is under the control of a single individual. It is created, used, and deleted by that person as required in that part of the business process for which he or she is responsible. Some examples of this type of data may be a spreadsheet, word processing document, or to-do lists. Clearly, personal data cannot be controlled or managed by the information systems group. Consequently, personal data is generally outside the scope of the data warehouse.

In the past, the majority of data of interest to an organization originated within that organization. The impact of external data on the organizational information architecture was relatively insignificant. This is no longer true. The growth of the Internet has caused an exponential growth in the volumes of external data entering an organization. External business data must be handled with great care when combining with existing internal data.

Data as a Product

This data is produced and stored for its own intrinsic value and not as a means of running or managing a business. It is a product of a business activity, can be

bought and sold, and must be managed and controlled like any physical product. For example, video and audio products such as movies and music recordings are increasingly produced, stored, and sold as digital data. Data as a product needs to be managed in a different way from business data, and is outside the scope of the data warehouse.

Metadata

Metadata describes the meaning and structure of business data. Business data is created, maintained, and accessed through business processes implemented with operational systems. Therefore, the business needs a full description of its business data and the processes by which to maintain and use it. Metadata describes a number of aspects of the business and of the corresponding application functions. It is an important component of the data warehouse because it is through metadata that data are registered, accessed, and controlled in the warehouse environment.

The metadata used in the process of defining and building business applications and their associated databases is *build-time metadata*. It facilitates consistency in use of data and functions. Build-time metadata is generated and stored in data modeling and application design tools such as computer-aided software engineering (CASE) tools. *Production-time metadata* is created to find, understand, and use the required data in the business. Just as operational data is the basic source for informational data, build-time metadata is the primary source of production-time metadata. Production-time metadata may be put to either active or passive use. Metadata that is used to control the action or function of some application or function has an active role. Metadata used in look-up mode to find some business data is being used in a passive mode.

Usage metadata is the most important type of metadata for the user of business data, particularly in the information environment. This is where the end user gains business benefit and improvements in productivity. Usage metadata describes the meaning of data and allows users to relate data elements or application function to their purpose in the business. It also expresses the relationship between the data (or application) and the organization responsible for maintaining the data.

Other types of metadata such as *currency metadata* and *utilization metadata* are actively used by the warehouse infrastructure as a mechanism to manage and control the operation of the warehouse. Currency metadata describes the timeliness of the business data. Whenever a change in data takes place that needs to be tracked, metadata is generated because metadata moves data from the

operational to informational level with a time stamp. For example, timing of a violation notice for environmental noncompliance is an important part of that event. Utilization metadata, on the other hand, is closely associated with the security and authorization functionality used to control access to the warehouse.

Notification data is another type of data associated with metadata. When data is entered into the data warehouse and into the metadata, a check is made to see who is interested in it. If the check shows that someone is interested in that information, a notification is sent to that person to alert him/her that data of interest has been captured. This type of function is particularly suitable for external data.

Granularity

Determining the level of granularity is the most important design issue in the data warehouse environment. Granularity refers to the level of detail or summarization in the data warehouse. The more detail there is, the lower the level of granularity. The less detail there is, the higher the level of granularity.

Granularity affects the volume of data that resides in the data warehouse and, at the same time, affects the type of query that can be answered. The volume of data in a warehouse is traded off against a query's level of detail. When data is used to manage a business, as is common in the data warehouse environment, it is rare that every event is examined. Taking a collective view of data is much more common. Using the high level of granularity is efficient if it contains sufficient detail.

The tradeoff in managing the issue of granularity of data must be considered very carefully at the design stage of the data warehouse. The best solution for most organizations is some form of multiple levels of granularity. Most DSS and EIS processing uses the compacted summarized or lightly summarized data. On those occasions where some greater level of detail is needed, however, there is the true archival level of data with multiple levels of granularity. Only when a data warehouse will contain a relatively small amount of data in its environment should a single level of data be attempted.

Partitioning

Another major design issue in the data warehouse is that of partitioning, which refers to breaking data into separate physical units that can be handled

independently. Having one large mass of data inhibits flexible access to particular data. Therefore, all current detail data is partitioned in the warehouse. The data may be divided by date, line of business, geography, organizational unit, or any other criterion. However, in the data warehouse environment it is almost mandatory that one of the criteria for partitioning be by date.

Data Structures

Many kinds of structures are found in the data warehouse. The simplest and most common data structure found is the *simple cumulative structure.* After daily transactions are transported from the operational environment, they are organized into data warehouse records by subject area and by date. All daily activity accumulated on a day-by-day basis is called simple cumulative data.

A variation of the simple cumulative data is the *rolling summary data*. Rolling summary data summarizes activity into seven daily slots for the first 7 days of the week. On the eighth day, the seven daily slots are combined and placed into the first weekly slot. At the end of the month, the weekly slots are combined and placed in the first monthly slot. At the end of the year, the monthly slots are combined, and the first yearly slot is loaded. Creating a continuous file from direct files is another data structure, and there are many more data structures within the data warehouse.

Distributed Data Warehouse

Most organizations build and maintain a single centralized data warehouse environment. At these organizations, most of the processing is done at a central headquarters, but a distributed data warehouse is needed in a few special cases. If much of the data is processed locally (or at a location apart from headquarters), some form of distributed data warehouse makes sense. Army installations have a great deal of autonomy, and a fair amount of processing that occurs at an Army installation's DPW. An installation DPW data warehouse is an example of a local distributed data warehouse. It is fed by its own operational systems and houses data unique to and of interest to the local operating site. This data warehouse contains data that is historical in nature and is integrated within the DPW.

A data warehouse can also be global. The global data warehouse studied for this report is for HQDA. Both global and local data warehouses contain historical

data. The global and local data warehouse contains data that is common across the corporation and data that is integrated.

Central to the success of the distributed data warehouse environment is the mapping of data from the local operational systems to the data structure of the global data warehouse. The global data warehouse at HQDA is designed and defined centrally so that each DPW data warehouse maps into the common structure. The detailed data resides at the local level, while the summarized data resides at the centralized global level. It is also possible to stage a global data warehouse at the local level, then pass it to the global data warehouse at the headquarters level.

Data can exist in either the local data warehouse or the global data warehouse, but not both. The minute redundant data exist between the local data warehouse and the global data warehouse, it indicates that the scope of the different warehouses has not been defined properly. When a difference of opinion occurs between the local and global scopes, it is only a matter of time before spider-web systems start to appear.

Underlying the whole issue of the distributed data warehouse is the issue of complexity. In a simple central data warehouse environment, roles and responsibilities are fairly straightforward. In a distributed data warehouse environment, however, the issues of scope, coordination, metadata, responsibilities, transfer of data, local mapping, etc. make the environment complex.

4 Data Warehouse Design and Development

Designing a data warehouse requires a number of techniques that address several design issues:

- *Enterprise data modeling* is a design technique that defines the contents of the warehouse to allow inclusion of the entire scope of the business.
- Since a data warehouse contains the historical data of the business, *techniques to structure and represent* historical data need to be considered during the design phase.
- A *common structure* is needed throughout the business data to allow end users flexible access to the data and to allow for combining data from multiple sources.
- The *strategic approach for populating the data warehouse* from multiple sources needs to be considered during the design phase as well. Proper design will enable an organization to minimize the ongoing maintenance problems of obtaining data from multiple, variable sources.
- At some point in time, data is purged from the warehouse. The issue of *purging data* is one of the fundamental design issues that must not escape the data warehouse designer.

Another design issue to consider is data transformation. To properly move data from the existing system's environment to the data warehouse environment, it must be integrated. Data extraction, transformation, and population into a data warehouse are considered to be the most technically challenging parts of building a data warehouse. The relationship between the source data and the target data drives data transformation. It is for this part of data warehousing that tools were first developed to try to reduce the effort involved. It still tends to be one of the most costly and time-consuming aspects of data warehouse implementation. Designing the data warehouse population approach, including the choice of replication tools, is one of the milestones in any data warehousing project.

Other issues to consider:

- End users need access to the information in the data warehouse in order for a business to realize the benefits. Data mart is a popular structure through which users gain access to the data warehouse.

- Archive and retrieval of the data in the warehouse should be considered at an early stage of design.
- To support the periodic nature of the data in the data warehouse, the designers must define the structure of the record timestamps.

Some other technical features also need to be considered during the development of a data warehouse. These technical features include the ability to manage a large amount of data, the database management systems, and managing external data. Several commercial tools are available to support different functions of building the data warehouse. All of these issues and techniques will be discussed more fully in the remainder of this chapter.

Data Modeling

Application Data Modeling

The design of the data warehouse begins with the data model, which is used to provide the user of the model with a clearer understanding of how the modeled objects behave. Application data modeling provides a logical view of the data required by the application, driven and defined by users' needs. It aims at developing specific business functions within the scope of a single application. This view is the basis for the logical and physical database design of that application. Application-level modeling provides a logical view of the data required by the application. It provides no significant support for integrating applications or for combining data from different sources. Supporting a combination of data from different sources requires a broader type of modeling, known as enterprise modeling.

Enterprise Modeling

Enterprise modeling directly supports the data warehouse and is usually the starting point of any data warehouse development effort. The focus of enterprise modeling is a complete and integrated view of all the data in the business. The data warehouse enterprise model is flexible and parallels the business structure rather than the data model of any particular application. The structure of the data in any single source application is inadequate for the data warehouse.

The most common forms of data modeling use the entity relationship approach. An entity is an object in which the business is interested. Each entity has a business definition, which is used to define the boundaries of the entity. Each

entity is associated with a number of attributes. An attribute is any characteristic of the entity that describes it.

The second major element of the entity relationship model (ERM) is the relationship. A relationship exists between the entities in a model and describes how the entities interact. Several modeling and diagramming techniques/tools can be used for the entity relationship approach. DOD and Army use IDEF* modeling techniques. See Figure 4 for an example of a partial IDEF model for DPW facilities management.

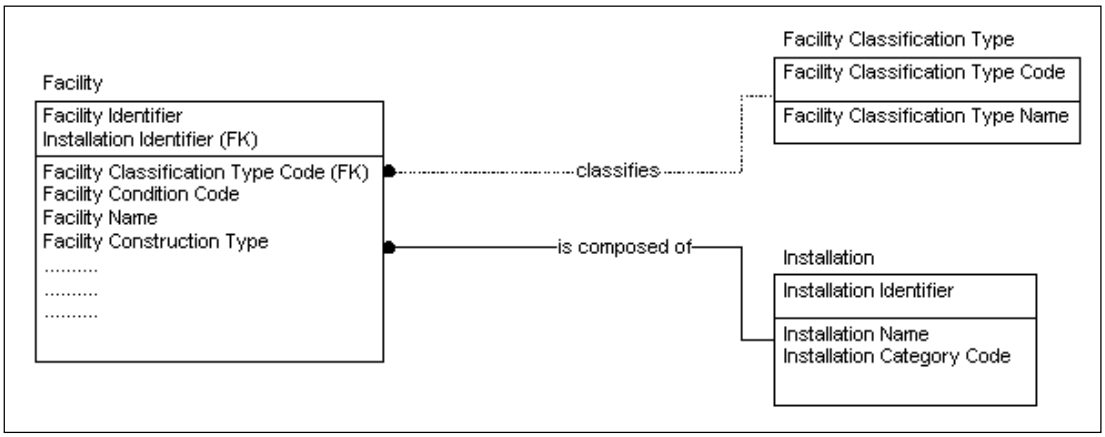


Figure 4. Example of a partial data model for DPW management.

"Facility" and "Installation" are entities. Their relationship is shown with a line, a verb on that line, and a dot. The attributes of facility and installation entities are listed in the boxes below the entity names. The facility entity is part of the installation entity.

Designing an Enterprise Model

The scope of a data warehouse covers the whole enterprise, and all parts of the organization must be involved in order to deliver a model that is valuable to the business as a whole. The approach, therefore, is to adopt a structure that allows for developing the model piece by piece, rather than as one large effort.

The first step is to acquire a highly consolidated view of the business to identify primary subject areas or concepts. Next, define the contents of the different

* IDEF = ICAM Definition; ICAM = Integrated Computer-Aided Manufacturing.

concepts in more detail. It helps to recognize the commonality between subject areas. For example, the relationship between a facility's M&R costs and condition of that facility's components and component structure needs to be recognized at a high level to ensure that data at lower levels in the model are correctly interrelated. Identification of commonality links the concepts to the generic ERM.

The next step is the ERM, which describes all data commonly used throughout the organization. A model developed in this way is enterprise-wide in its scope and is generic to all of the application views. DOD is working on an enterprise model for DOD as an enterprise. Army is developing an HQDA data warehouse with HQDA as an enterprise. An ERM to cover the overall DPW functions can be considered a DPW enterprise model. The DPW enterprise model needs to be consistent with these DOD enterprise and HQDA data warehouse models. Some of the entities will be used almost exclusively by one part of the organization, but such localized entities can be defined separately from the overall company-wide aspects of the model. This compartmentalization provides a practical approach to implementation, allowing the model to be defined in stages.

Application data models are closely related in content to the enterprise-wide generic model. A single entity in the generic ERM can appear a number of times in the logical application views, with different attributes, in order to meet the needs of different business applications. It is the relationship between multiple entities in this layer and a single entity in the layer above that ensures that the resulting applications use data consistently. It also indicates shared data between applications.

Data definitions created as part of enterprise modeling form part of the metadata in the warehouse. By providing an efficient and user-friendly means of accessing and using this metadata, the data warehouse ensures that end users can obtain business benefit from the warehouse.

Corporate Data Modeling

The corporate data model is used for the design of the operational environment, and a variation of the corporate data model is used for the data warehouse. A fair number of changes, however, are made to the corporate data model as it is applied to the data warehouse. The first change is to remove data that is used purely in the operational environment. Next, the key structures of the corporate data model are enhanced with an element of time. Derived data is also added to the corporate data model where the derived data is publicly used. The final design activity is to perform stability analysis. Stability analysis is the act of grouping attributes of data together based on their propensity for change. Data

that seldom change are grouped with other data that seldom change, data that sometimes change are grouped with other data that sometimes change, and data that frequently change are grouped with other data that frequently change. The net result of stability analysis is to create groups of data with similar characteristics.

High-level modeling at entity relationship level is called ERD (for entity relationship diagram). The scope of integration defines the boundaries of the data model and determines what entities belong in the model. The scope of integration must be defined before the modeling process commences. If the scope is not predetermined, chances are high that the modeling process will continue forever. The corporate ERD comprises many individual ERDs that reflect the different views across the corporation.

The next level is the *mid-level model*. For each major subject area, or entity, identified in the high-level data model, a mid-level model is created. The mid-level model defines primary and secondary groupings of data with attributes, relationships of data between major subject areas, and type of data. These data modeling constructs are used to identify the data attributes in a data model and the relationship between those attributes.

The *low level* or *physical data model* is created from the mid-level data model by extending it to include keys and physical characteristics of the model. At this level, the physical data model looks like a series of tables, sometimes called relational tables. Performance characteristics need to be factored in while designing the physical database. Granularity and partitioning of the data, physical input/output (I/O), and other design activities are included in the design to ensure good performance results out of the data warehouse environment. The physical data model should allow the different iterations of development to be built in a cohesive manner. When the different iterations of development are done without a unifying data model, there is much overlap of effort and much separate and disjointed development.

The output of the physical data model process is a series of tables, each of which contains keys and attributes. When there are many tables, a large quantity of I/O resources are required. Some of these tables can be merged to minimize use of I/O resources. Merging tables, creating an array of data, selective use of redundancy, derived data, creative indexes, and artifacts in the data warehouse are some of the normalization/denormalization techniques used in the data warehouse design phase to save I/O and improve performance.

Historical Data

Because a business changes over time, the business data must represent that change. A data warehouse must explicitly consider the temporal aspects of the data it contains because it must, by definition, provide a historical view of the business. The techniques for representing the time dimension of data lead to the ability to store historical data in a way that supports its use for many business purposes throughout the organization.

Several approaches represent the time dimension of data. One widely used method is the application of timestamps to the data. Timestamps may be applied at a field level, record level, or even at the file/table level. For example, whenever any field in a record is changed, the record timestamp is updated wherever it has been applied. It is often up to end users to decide on what level they wish to track the currency of their data.

Another issue for recording historical data is how to capture the changes in the data and represent it over time. Data are changed through business transactions. The status of the data at a given time is critical for most applications. The status-based database stores a large volume of data, because every change to the status of the data updates it and, in the case of a data warehouse, duplicates most of it. The amount of data is likely to be less when an event approach is used. It updates only primary key and changed fields. Timestamps, together with the concepts of status or event representations, allow the maintenance of temporal data.

In periodic data, once a record is added to the store, it is never physically deleted, nor is its business content ever physically modified. Rather, new records are added, even for updates to or deletions of existing records. Periodic data thus contains a complete record of the changes that have occurred in the data. Either statuses or events can form the basis for this complete record. Accessing historical data is one of the primary incentives for adopting the warehousing approach, which uses historical data for trend analysis, prediction, and discovery of patterns in a particular area of the business data.

Probably the most obvious concern regarding historical data is its potential volume and the associated costs of storing it. However, the volume of historical data that should be retained must be considered in terms of its potential business benefits. If all data are stored at the highest level of detail and never deleted, then all possible future queries and analyses can be supported. However, this approach can be difficult to justify by a cost-benefit analysis. Summary data is generally used over a longer time span than detailed data. Such an analysis of

the need for and use of historical data is carried out at the data warehouse design phase. The likelihood of future data requirements, as well as the consequences of not having that data, must be weighed against the cost of storing and managing that historical data.

Data Mapping, Extraction, and Transformation

Data transfer within and between different systems requires a set of techniques that provide comprehensive support for copying and transforming data from source to target location in a managed, consistent, repeatable, and well-understood manner. Data transfer as it relates to the data warehouse exhibits certain characteristics — consistency of data, reuse capabilities, integration of metadata, and ease of maintenance.

While the data architecture and modeling activities constitute a significant challenge in the design phase of the warehouse, the implementation of the population function is often the most costly and time-consuming part of the entire implementation. The physical locations of source and target data can significantly limit the choice of tools for implementing data transformation. Trade-offs are made between the depth of a function, breadth of platforms supported, ease of maintenance, flexibility to support changing business data needs, and performance. The process of data transfer and the steps in the process need to be well-defined during the data warehouse design phase:

- Identify source and target data requirements through data models
- Create the mapping between source and target
- Determine if the data transfer from source to target is going to be a bulk transfer or transfer of only changed data
- Define the schedule of data transfer at specified intervals
- Determine where the data transformation is best performed
- Transfer the data between source and target based on the defined mapping, and document.

Establishing the relationship or mapping between the source and target data is the first and most important requirement in data transfer strategy. Several details have to be programmed just to bring the data from the operational environment properly. Data may not be encoded consistently, the field may be measured differently in different applications, or the same field may exist in different applications in different names. To transform the data to the data warehouse properly, there must be a mapping from the different source fields to the data warehouse fields. The issue of integrating data can be complex and burdensome.

Integration of existing systems is not, however, the only difficulty in the transformation of data from the existing system's operational environment to the data warehouse environment. Another major problem is that of the efficiency of accessing the existing system's data. Usually the existing system's environment can be downloaded to a sequential file and the sequential file can be downloaded into the warehouse with no disruption to the online environment. Still the loading of data on an ongoing basis presents the largest challenge to the data architect. Efficiently trapping those changes and manipulating them is not easy. A determination has to be made on how often and how much data to scan. A shift in the time basis is also required as data is moved from the operational to the data warehouse environment.

Finally, condensation of data is a major consideration when data are passed from the existing system's environment to the data warehouse environment. The data may have to be condensed from daily to weekly, monthly, or even yearly. Condensation of data is vital in warehouse data management. If the volumes of data are not carefully managed and condensed, the sheer volume of data that aggregates in the data warehouse prevents the goals of the warehouse from being achieved.

Star Joins

The "star join" design structure is required to manage large quantities of data residing in an entity in a data warehouse. A simple two-dimensional data model gives the impression that all entities are equal. Some entities are sparsely populated, however, while other entities are heavily populated. Figure 5 shows such a three-dimensional perspective. A work order entity, in the case of DPW management, will have many more occurrences of data residing in the table(s) than the installation entity or facility entity. Work order, the heavily populated entity, is at the center of the star join. Other, less populated entities are the surrounding entities. The center of the star join is the "fact" table. Surrounding entities are "dimension" tables. The fact table contains unique identifying data and foreign key references that are (prejoined) to the dimension tables.

Creating star joins streamlines data for DSS processing. By prejoining data and creating selective redundancy, the designer greatly simplifies and streamlines data for access and analysis, which is exactly what is needed for the data warehouse. Data modeling applies to the dimension tables and star join design applies to the fact tables.

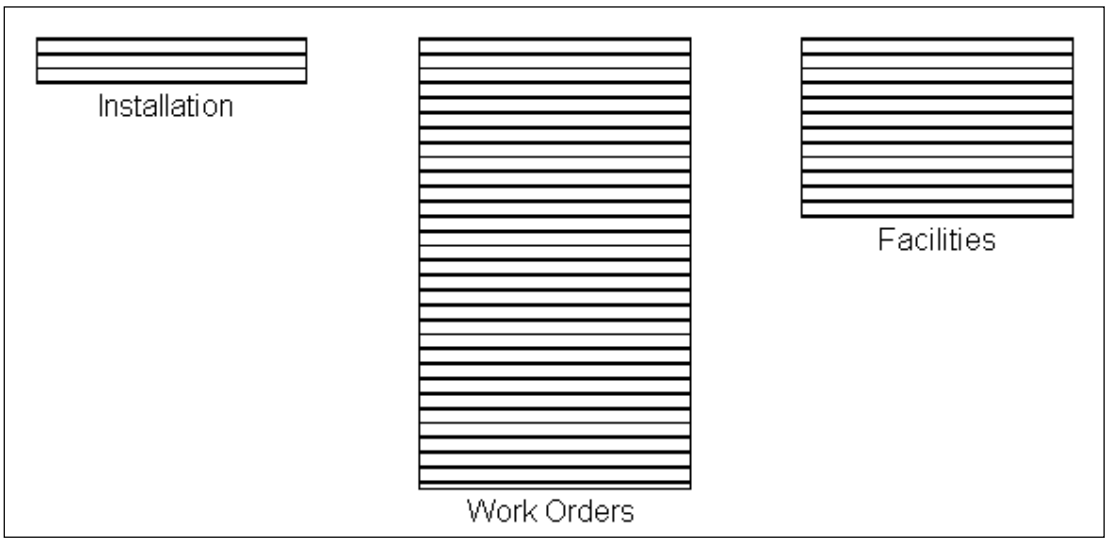


Figure 5. Work orders fact table with occurrences of data in a star join.

Data Partitioning by Date

As discussed in Chapter 3, other issues to consider during the design of a data warehouse are granularity and partitioning. When the granularity of a data warehouse is properly set, the remaining aspects of design and implementation flow smoothly. Partitioning of data into separate physical units allows each unit to be handled independently. How it is partitioned is up to the developer. However, in the data warehouse environment it is almost mandatory that one of the criteria for partitioning be by date. One of the essences of the data warehouse is the flexible access of data. Having a big mass of data defeats much of the purpose of the data warehouse. Therefore, all current detail data warehouse data will be partitioned.

Database Management System (DBMS)

General purpose DBMS products can be used for the data warehouse, DSS, and EIS. With the advent of data warehousing and new focus on DSS, a new class of DBMS has arisen. The new class is called data warehouse-specific DBMS products. A warehouse-specific DBMS product is optimized for data warehousing and DSS processing. End users who need access to large volumes of data stored in a general-purpose DBMS are often frustrated by poor response times and lack of flexibility. Warehouse-specific DBMS products provide better performance and flexible retrieval.

The advantage of a general-purpose relational product is that it follows industry standards and is familiar to most persons in the organization. Also, warehouse-specific DBMS products are commonly implemented with proprietary DBMS products, which is a problem. There is considerable industry debate concerning the appropriate choice of DBMS for the data warehouse environment. Hyperion's Essbase® is a multidimensional warehouse-specific DBMS. Red Brick Warehouse by Red Brick® Systems, Inc. is a relational database designed specifically for data warehouse. Oracle, a general-purpose DBMS, claims that Oracle7 provides exceptional data warehouse performance through advanced optimization techniques, parallelization, scalability, and join methods. Multidimensional server vendors point out that their products complement, but do not compete with relational databases. Oracle is also working on extensions to enable Oracle users to perform multidimensional analysis.

Data Warehouse Development Tools

Many vendors are manufacturing hardware, software, and tools to help data warehouses function effectively. There are tools for data modeling, extraction, cleaning, loading, storage, and mining. IDEF1X data models, developed using the ERwin product by Logic Works are DOD- and Army-approved.

Informatica Corp.'s PowerMart suite contains Informatica PowerMart Designer, Repository, Server Manager, PowerMart Server, and Change/Capture components. Popular relational databases are supported. The Star Schema Design Wizard uses a visual interface to step a user through the database design. Once completed, the SQL data definition language is generated for the target database. The Change/Capture function accesses operational system log records to move incremental change records into the data mart. Changes can be captured as they occur or on a periodic basis. Table 1 lists popular databases suitable for data warehousing.

The tasks of extracting, cleaning, and loading information into a data warehouse take an enormous amount of time. Inmon (1996) estimates that, on average, 80 percent of the efforts of building a data warehouse go into these tasks. Tables 2, 3, and 4 give a list of some popular data extraction, cleaning, and loading tools.

Most data extraction tools provide data loading capability also. They automate parts of the extract and load tasks.

Table 1. Databases for data warehousing.

Vendor	Product	Phone
Arbor Software Co.	Essbase	800-858-1666
Computer Associates	OpenIngres	800-225-5224
Dimensional insight Inc.	CrossTarget	617-229-9111
IBM	DB2	800-426-2255
Kenan Technologies	Acumate Enterprise	617-225-2224
Microsoft	Microsoft SQL server	206-635-7041
Micro Data Base Systems	Titanium	317-463-7200
NCR	Teradata DBS	513-445-5000
Oracle	Oracle	800-633-1071
Oracle	Oracle server	800-633-1071
Pilot Software Inc.	LightShip	617-374-9400
Red Brick Systems	Red Brick Warehouse	408-399-3200
SAS	SAS System	919-677-8000
Software AG	Adabas D	800-423-2227
Sybase	Sybase IQ	510-922-3500
Sybase	Sybase SQL server	510-922-3500
VMARK	UniVerse	508-366-3888
XDB Systems	XDB Server	800-488-4948

Table 2. Data extraction tools.

Vendor	Product	Phone
Alpha Microsystems	AlphaConnect	714-957-8500
Carleton	Passport	617-272-4310
Computer Associates	CA-LDM	516-342-5224
ETI	ETI-Extract	512-327-6994
Gladstone	Gladstone data package	800-709-7800
IBM	DataPropagator	800-426-2255
IBM	Visual Warehouse	800-426-2255
Informatica	PowerMart	800-653-3871
Information Builders	Enterprise copy manager	800-969-4636
Kapstone Systems	Thazar	816-760-5000
Prism Solutions	Warehouse manager	800-995-2928
ReGenisys	Rule Finder	800-401-7853
SAS Institute	SAS Data Warehouse	919-677-8000
Software AG	SourcePoint	800-694-4111
Syware	DataSync	617-497-1376
Data Junction Corp.	DJXL	800-580-4411

Table 3. Data cleaning tools.

Vendor	Product	Phone
Applied Parallel Technologies	Orchestrate	617-494-1177
Gladstone computer services	Gladstone data package	800-709-7800
Harte-Hanks data technologies	Trillium software system	508-663-9955
Innovative systems	Innovative Warehouse	800-622-6390
Mercantile software systems	IRE Marketing warehouse	908-981-1290
Platinum Technology	InfoRefiner	708-620-5000
Postalsoft	Postalsoft Library	800-831-6245
QDB Solutions	QDB/Analyze	617-577-9205
Sagent Technology	Sagent Data Mart solution	415-833-6800
SAS Institute	SAS Data Warehouse	919-677-8000
System Techniques	Converge Tool Set	404-814-3850
Vality Technology	Integrity Data Reengineering	617-388-0300
Vmark	DataStage	800-966-9875

Some special purpose tools are also available. Advanced Technologies has a product to load text files into personal computer (PC) databases. AutoImport by White Crane Systems provides the means to extract data from information stored in report files. Dataflux Corporation has software that cleans name and address information. DataCleanser by EDD, Inc. assists in the cleaning of Informix databases. Some tool sets automate several tasks. Converge tool set, Powermart®, DJXL, DataStage, Sagent, and SAS data warehouse are some of the products that can automate parts of the extract, clean, and load tasks. Table 4 lists some popular data loading tools. Some of the newer integrated data warehousing tool sets include tools for metadata management, scheduling, data warehouse organization, and data transformation.

Table 4. Data loading tools.

Vendor	Product	Phone
Applied Parallel Technologies	Orchestrate	617-494-1177
IBM	Data Propagator	800-426-2255
Mercantile software systems	IRE marketing warehouse	908-981-1290
Platinum Technology	Info Transport, Fast load	708-620-5000
Praxis International	OmniLoader	508-270-6666
SAS Institute	SAS Data Warehouse	919-677-8000
Smart Corporation	Smart DB Workbench	415-988-8996
Spalding Software	DataImport	770-449-0594

Data Mart / Data Store

Data marts are user-community-specific data stores that focus on decision support system end-user requirements. This approach is conceptually a traditional approach, delivering specific data to groups of users as required. The difference is that data mart provides an information systems structure that allows an organization to have very flexible access to data, to "slice and dice"* data any number of ways, and to dynamically explore the relationship between summary and detail data. The data mart database focuses on one subject area, while an enterprise data warehouse contains information from many different subject areas.

Industry experts disagree on whether data warehouse and data mart are two mutually exclusive architectural alternatives for enterprise-wide decision support. Some believe that organizations should blend these two approaches in a multitiered strategy (Inmon 1996; Demarest 1994). Their approach is to combine data warehousing and data marting, resulting in the enterprise DSS. The vendors of OLAP servers and multidimensional databases argue that relational databases are incapable of efficiently meeting end-user requirements and propose specially optimized software. Vendors of relational databases and decision support tools that operate against such databases point to star schemas, parallel processing, and improved indexing technologies to handle relational data efficiently.

Vendors of multidimensional database and analysis tools tout multidimensional technology as database technology. They claim that relational databases are not well-suited to flexible data analysis so multidimensional database engines and analytical tools should be used for data mart and data warehouse development and analysis. Inmon and some vendors of relational databases argue that multidimensional DBMS technology is not database technology. They recommend using multidimensional products as complementary to relational products, but not instead of relational DBMS. However, data marts developed for specific users, once in production, are difficult to extend for use by other departments, if the data definitions, timeframes, and data names are not consistent with applications in other departments. Also, this practice would be contrary to the data warehouse approach, which has a single source of data.

* The term "slice and dice" is more fully explained on page 54.

Two basic techniques used to build a data warehouse are known as the “top down” and the “bottom up” approaches. In the “top down” approach, a data warehouse is first built for the complete organization. This will be a huge database where all the end-user information is stored. From this database, information needed for local end users can be selected. In the “bottom up” approach, data marts — smaller local data warehouses — are accessed by end users at a local level for specific local requirements. The advantage of the bottom up approach is that a local data warehouse can be built in a very short time and can be managed at a departmental level, with each data mart completely optimized for particular tasks. For this reason, numerous data marts are often found in an organization. The enterprise data warehouse is generated out of these. However, organizations need to be able to centrally administer and manage these databases to prevent the proliferation of data marts that contain inconsistent and conflicting data — precisely the problem that data warehousing is supposed to solve.

Several fundamental approaches can give users access to decision-support data. One approach is to build an enterprise data warehouse that can be used directly by users. A second approach is to build data marts planned for eventual integration into a data warehouse. Another technique that is becoming popular is to build the infrastructure for an enterprise data warehouse while at the same time building one or more data marts to satisfy immediate business needs.

In a data warehousing approach, data flows from the data warehouse into the multidimensional DBMS. An organization can have one enterprise data warehouse and many data marts. Generally, data marts contain less information than the data warehouse and are more easily understood and navigated than enterprise data warehouses. A data mart will provide better performance and faster user response than an enterprise data warehouse. As data marts grow in size, however, performance may worsen. Data marts (sometimes called multidimensional DBMS) provide an information systems structure that allows an organization to have very flexible access to data. Data mart provides a capability to “slice and dice” data any number of ways and to dynamically explore the relationship between summary and detail data.

The detailed data housed in a data warehouse provides a very robust and convenient source of data for the data mart. Data flows from the data warehouse into the multidimensional DBMS on a regular basis. Since operational data is integrated as it enters the data warehouse, the multidimensional DBMS benefits by not having to extract and integrate the data. The data warehouse acts as a “wholesale” source of data, and that data is “retailed” to the data marts based on local need. Data marts provide flexibility and control to the end user.

The final point to consider is how a data mart fits into the architecture of a data warehouse. There is a complementary relationship between a data warehouse and a data mart. For example, a data warehouse may be designed as a relational database, while a data mart is designed for multidimensional analysis. The size of data, the number of users, and the type of analyses that are performed determine the architecture and technical implementation methods.

5 Decision Support and the Data Warehouse

Managers use experience and knowledge to make good decisions every day. For decisions involving lots of information, possible conflict, and a big commitment, however, conventional data-processing reports are not the best decisionmaking tools. These reports proceed in a linear fashion, presenting a single view of the data. This view may be partial or misleading. Multiple analysis procedures on consolidated and integrated information in a data warehouse use a goal-directed framework that serves the needs of management in the decisionmaking process. One of the main reasons to build a data warehouse is to provide a framework for DSS processing and to support executive decisionmaking. The problem today is not shortage of data. The real problem is to make sense of the data and arrive at the best decision.

To make a well-informed decision, management must gather and analyze a large quantity of information. For example, if the goal is to select a site for a pipeline routing, management makes a decision based on an analysis of the alternative sites for costs, operational efficiencies, and environmental and socioeconomic considerations. Consolidated information is necessary to perform such analysis. The information may be in Relational DBMS (RDBMS) format, spreadsheet format, textual information, geographic information systems (GIS) format, computer-assisted design (CAD) format, or images.

To assist managers in making the right choices, it is essential to be able to research the past and identify relevant trends. Setting up a data warehouse is the easiest way to gain access to all types of information with historical data from which to facilitate effective decisionmaking.

The RDBMS is an excellent tool for organizing large amounts of data and for defining relationships between data sets in a consistent way. Although there are ways to manage multidimensional data in an RDBMS, most systems do not provide tools to view and analyze the data in a multidimensional format. In recent years there has been a growing awareness that a great deal of business data is multidimensional in nature. Conventional tools are not optimized for multidimensional data; they therefore present an obstacle to rapid and accurate

analysis. A number of specialized tools have been developed to address these concerns. These OLAP tools store data in a multidimensional form, and allow the user to look at segments of the multidimensional cube or to "drill down" through summary-level data to detail data. OLAP tools will be discussed further in the next section.

DSS tend to focus more on detail and are targeted towards lower to mid-level managers. EIS have generally provided a higher level of consolidation and a multidimensional view of the data, as high-level executives need to "slice and dice" the same data more than they need to drill down to review the data detail. DSS and EIS have data in descriptive business terms, and these systems are designed for use by nontechnical users. The data are generally preprocessed with the application of standard business rules, and consolidated views of the data are available. Data warehousing systems are most successful when their design aligns with the overall business structure rather than specific requirements.

As databases have grown, there has been an accompanying explosion in the potential for data mining. Data warehouses tend to contain extremely large data sets. Data mining deals with the discovery of hidden knowledge, unexpected patterns, and new rules from these large databases. It involves the use of machine learning, statistics, or knowledge discovery techniques such as rule induction, classification, clustering, pattern recognition, predictive modeling, and dependency detection. Mining operational data is almost impossible because there are different applications with different types of attributes and different data types but no historical data. With a data warehouse, this problem does not exist — all the information has been cleaned, integrated, consolidated, and transferred.

OLAP / Multidimensional DBMS

Relational tables competently describe the what, when, where, who, and how, but they often cannot answer why things happened the way they did. Answering this question usually requires more complex analysis. Multidimensional database engines and analytical tools optimize data storage and manipulation for output to help users investigate patterns not easily revealed by transactional reports. Multidimensional analysis (or OLAP) is an analytical technique that allows users to view their data in a dimensional cube format, and to easily select and analyze that data. Multidimensional analysis allows end users without extensive mathematical or statistical training to perform operations such as drill-down, roll-up, cross tabulations, ratios and trends, slice and dice, and data pivoting. Some of the results needed by the decisionmaker may not be stored in the

data warehouse, but are calculated dynamically from warehouse data in response to each request.

A well-implemented multidimensional database allows the end user to quickly focus on the exact view of the data required. The end user may want to determine, for example, how the maintenance costs of metal roofs compare to maintenance costs of asphalt shingle roofs. A manager may want to compare maintenance costs on buildings with maintenance costs on site- and landscape-related activities. The end user selects the desired positions along each dimension. OLAP also enables the creation of hierarchies within each dimension. For end users, the ability to define hierarchies allows for very quick data manipulations and detailed analysis along different levels within all the dimensions of a multidimensional array.

OLAP tools support the drill-down and roll-up processes as long as the data that are needed are available and structured properly. The OLAP servers are almost exclusively used to build data marts. Though it is technically possible for the OLAP data tables to reside in a data warehouse database, it is not recommended. These products are optimized to suit the needs of an interacting workgroup rather than serve as a component of a more generalized access and analysis infrastructure. They support the autonomy of the functional unit by serving the data the way they need it, on a platform they control, with functionality to help them in their work. These products produce standalone islands of data like their predecessors, but offer a far richer array of access, analytic, and interactive capabilities. What keeps them from being isolated islands is a well-crafted data warehouse environment that acquires the detail data and maintains the history that will provide the basic raw data these tools will draw upon.

The relationship between OLAP and data warehouse is interesting and complementary. The detailed data housed in a data warehouse provide a very robust and convenient source of data for OLAP. If data warehouse and OLAP are designed properly, the data mart with OLAP can store all but the most detailed level of data. Once in the multidimensional format, the data can be further summarized. The analyst using the OLAP tools can drill down in a flexible and efficient fashion over all the different levels of data found in the data mart. Then, if needed, the analyst can actually drill down to the data warehouse. Another complementary aspect of the data warehouse coupled with OLAP is that OLAP data cover short lengths of time depending on the application, whereas a data warehouse spans a much longer time horizon. In this way, the data warehouse becomes a source of data for OLAP analysts.

Several vendors provide OLAP products. Some products access data stored in relational tables, while others access data stored in multidimensional databases. Some popular OLAP products are listed in Table 5.

Some of these tools provide the ability to perform OLAP analysis over the Internet or a corporate intranet. Some tools store their data in multidimensional format, while others store in relational format and present multidimensional views. OLAP tools are developed to answer any question about the data and allow the user to quickly summarize selected information.

Table 5. Online analytical processing (OLAP) tools.

Product	Vendor	Phone
Arbor Essbase Analysis Server	Arbor Software Co.	800-858-1666
BrioQuery	Brio Technology	800-879-2746
Business Objects	Business Objects, Inc.	800-705-1515
DB2 OLAP Server	IBM	800-426-2255
Decision	Comshare	800-922-7979
Decision Suite	Information Advantage	800-959-6527
DSS Agent	MicroStrategy	800-927-1868
Express Server, Objects	Oracle	617-768-5600
Fusion	Information Builders	800-969-4636
Gentia	Gentia	303-794-8701
Holos	Seagate Software	908-321-6500
Hyperion MBA	Hyperion Software	800-286-8000
Metacube	Informix	415-288-7966
Pilot Decision Support Suite	Pilot Software, Inc.	800-944-0094
PowerPlay	Cognos Corporation	800-426-4667
SAS System	SAS	919-677-8000

Executive Information Systems

Through EIS the executive analyst can pinpoint problems and detect trends that are of vital importance to management by running trend analyses and comparisons. The EIS analysis alerts the executive as to what the trends are. It is then up to him or her to discover the underlying reasons for the trends. Many EIS are developed using OLAP tools.

After running a trend analysis (Figure 6), an executive has isolated good, fair, and poor training facilities over several years. Looking just at poor facilities, the executive identifies a trend. Each year the condition of the training facilities is falling. Having identified the trend, the executive can further investigate why training facilities are falling to poor condition.

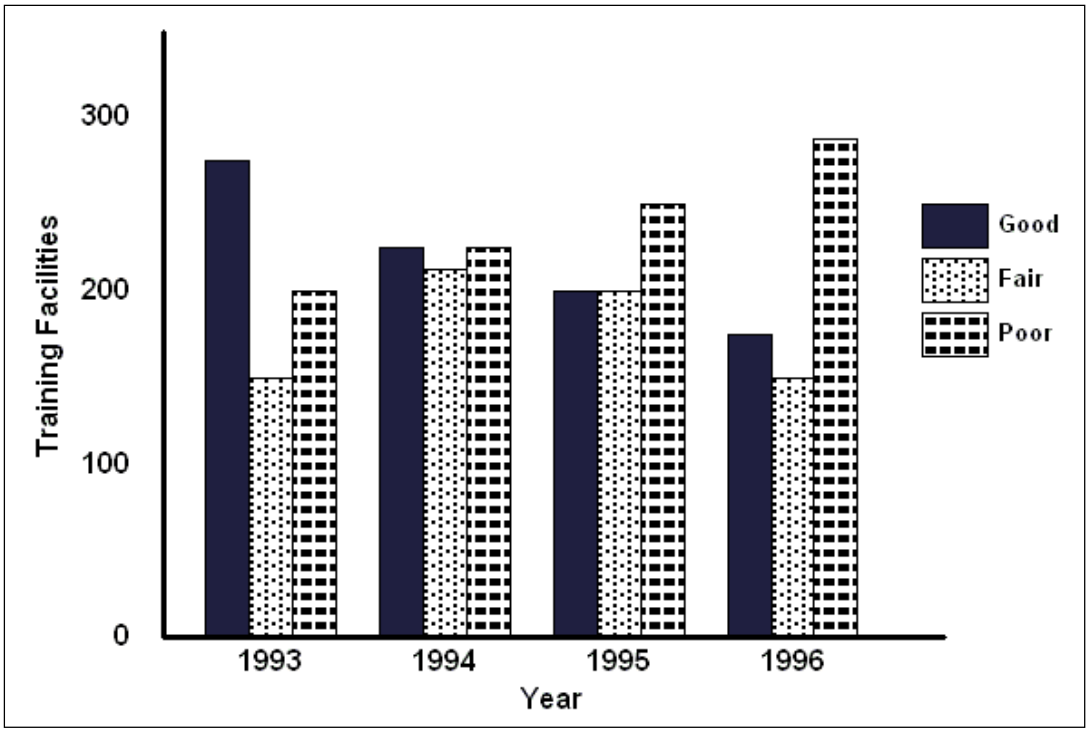


Figure 6. A typical EIS processing chart.

Comparisons are another type of useful analysis. Figure 7 shows a comparison that might be found in an EIS analysis. Looking at 1997 M&R costs and 1996 M&R costs, the question can be asked, why is there such a difference in costs between those 2 years? The EIS processing alerts the manager to these differences. It is up to the manager to determine the underlying reasons.

Trend analysis and comparisons are not the only methods accommodated by EIS. In the "slice-and-dice" approach, the analyst takes basic information, groups it one way, and analyzes it; then it is grouped another way and analyzed again. Slicing and dicing allows the manager to have many different perspectives. In order to slice and dice, it is necessary to be able to drill down data, by starting at a summary level and breaking that summary data into detail.

Figure 8 shows that the manager wants to explore M&R costs further. The manager looks at the types of accounts that have contributed to the growth in M&R costs in 1997. In looking at the numbers for each account, the manager decides to look at the K account more closely by facility type. Of these facility types, the manager then decides to look more closely at the numbers for buildings, and then at major components for buildings. In each case, a path going from summary to detail is determined. In such a fashion, the manager can determine where the troublesome results are.

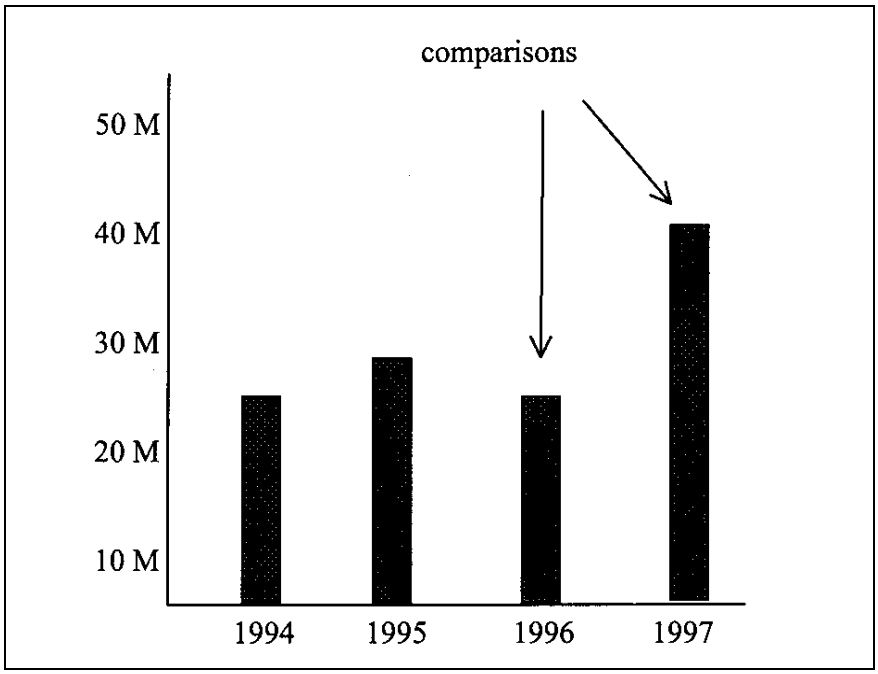


Figure 7. Comparing 1997 M&R costs to 1996 M&R costs.

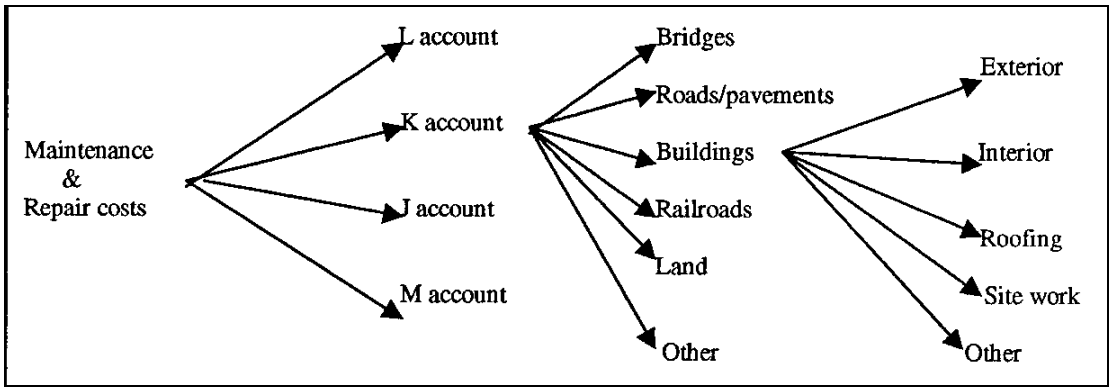


Figure 8. Example for drill-down analysis.

EIS software supports the drill-down process as long as the needed data are available and structured properly. Management's focus shifts with every new problem or opportunity that arises, so the EIS needs to have the data required for different types of analysis. The data warehouse is tailor-made for the needs of the EIS analyst. Once the data warehouse has been built, the job of the EIS is made infinitely easier with a foundation of data on which to operate.

Decision Support Systems

The sole function of a data warehouse is to supply the information needed to make knowledgeable decisions. In some cases, standard SQL tools may be used for decision support. For comparing millions of records without knowing the

exact type of information required, or to find hidden data, then data mining will serve as a decision support tool. Some end users are interested in only a particular part of the information with trend analysis, risk analysis, and other decision support methods using statistical techniques. All these types of functions support management in making decisions.

Most organizational problems involve more uncertainties than facts. DSS helps quantify these uncertainties and assists managers in reaching decisions by applying a rational, step-by-step structure. In such an approach, the problem is broken down into a logical progression of specific actions. The process is usually represented as a decision tree. Each factor in the decision tree is assigned a cost or probability by the decisionmaker. This structure gives the decisionmaker an efficient way to compare alternatives. Management is able to change the weighting or the criteria as it sees fit, and to recalculate the tree to determine what the decision would be under new criteria.

For example, requirements forecasting is the baseline for Army installation master plans. The key to the overall forecast is the character of the basic operation, total requirements, and staged development of the installation. Basic operations depend on missions and their characteristics. Total requirements include land, infrastructure, facilities, and services requirements (e.g., sustainment, improvement, acquisition, diversion/conversion, and disposal requirements). The staged development depends on mission requirements and criticality, condition of the existing facilities, costs of development, and availability of alternatives. Several decisions must be made at every step of forecasting and the master planning process. Decision support tools assist a master planner by providing adequate information to support these decisions.

The Army Real Property Planning and Analysis System (RPLANS) is a DSS — it helps planners to analyze real property requirements. The economic analysis package for military construction, ECONPACK, is another DSS — it provides information on the cost benefit analysis which is a major factor in decisionmaking. RMAT, a real property management tool, is also a DSS — the objectives of RMAT are to analyze the carrying capacity and to determine capital investment strategy. There can be several decision support tools for different objectives, strategies, techniques, and decisions. If the underlying data for these multiple decision support tools are not consistent, however, the analysis will not be accurate.

A data warehouse can be the single source of data for all these DSS. The current approach is to develop interfaces between systems. This approach creates spider-web architecture, which has already been shown to be undesirable. Also,

most of the existing systems do not contain historical information, which is necessary for most analysis. A data warehouse provides historical data that can be a rich source of information for trend analysis, forecasting, and risk analysis. The data warehouse architecture with integrated information can be the single source of data for all DSS.

Data Mining

Modern organizations are under enormous pressure to respond quickly to changes in the market. To do this they need rapid access to all kinds of information before they can make any logical decisions. Most organizations have large databases that contain a wealth of information. However, it is usually very difficult to access this information. Some knowledge hidden in the databases is hard to find using SQL. This difficulty in finding desired information in huge databases has led to a growing interest in the field of data mining.

Data mining finds answers to questions that analysts have not thought to ask. It discovers information within data warehouses that queries and reports cannot effectively reveal. The potential payoffs from data mining are enormous if the right tools are chosen and used effectively. These applications can become the foundation of an organization's strategies. Traditional database queries are designed to supply answers to simple questions. OLAP lets users run much more complex queries. In both cases, however, the results are merely extracted values or an aggregation of values. Data mining reaches much more deeply into databases. Data mining tools find patterns in the data and infer rules from them. These patterns and rules can be used to guide decisionmaking.

Although it is not strictly necessary to have access to a data warehouse in order to carry out data mining successfully, in practice it helps quite a bit. Obviously, in order to perform any trend analysis requires access to all the information needed to support that analysis, and this information is stored mainly in large databases. The easiest way to gain access to this data and facilitate effective decisionmaking is to set up a data warehouse.

Because there are several types of data mining techniques, it is important to understand the demands of the end user so that a proper data warehouse is built for data mining. Any technique that helps extract more out of data is useful, so data mining techniques are quite a heterogeneous group. Some popular techniques are: statistical, classification, clustering, rule induction, decision tree, k-nearest neighbor, and neural network. These techniques will be described in the remainder of this section.

Data is a strategic asset. Its effective use can provide information that can change the way a business is managed. Intelligent data mining requires a mix of tools and techniques. Users need to know both the tools and the business.

The first step in a data mining project should always be a rough analysis of the data set using traditional query tools. Before more advanced pattern analysis algorithms can be applied, some basic aspects and structures of the data set need to be known. Some interesting information can be abstracted from a database using SQL, but hidden knowledge cannot be found without more advanced data mining techniques.

One of the earliest statistical methods used in data analysis is multiple regression, a standard statistical technique that uncovers the pattern of dependencies between multiple predictor fields and the outcome. Associations, sequences, and forecasting are some other methods based on statistical techniques. *Associations* happen when occurrences are linked in a single event. For example, a study of work orders might reveal that, when a roof is replaced, 60 percent of the time exterior is also replaced. In *sequences*, events are linked over time. If interior is repaired, 55 percent of the time furniture will be replaced within 1 month. *Forecasting* estimates the future value of continuous variables — like service orders — based on patterns within the data.

Classification techniques recognize patterns that describe the group to which an item belongs. It does this by examining existing items that have already been classified and inferring a set of rules. This activity is probably the most common in data mining today. For example, classification can help discover the characteristics of projects that are likely to go over estimates and provide a model that can be used to predict them. It can also help determine which kinds of service orders are likely to need repetitive work, so that managers can schedule and manage work effectively.

Clustering is similar to classification, but differs in that no groups have yet been defined. Using clustering, the data mining tool discovers different groupings within the data. This technique can be applied to problems as diverse as detecting defects in equipment or finding suitable groups of people for self-help projects.

Rule induction generates rules based on some conditions, associations, attributes, processes, or calculations. These applications may involve predictions, such as whether a customer will use a self-help project.

A decision tree is a graphical representation of a set of rules that classify data. These decision trees and rules can be complex based on variables and classes, yet decision trees are easily understood. They show the combined dependencies between multiple predictors and the outcome as a number of decision branches. An analyst can more easily control the decision tree model construction and can assemble a more valid and reliable final result. Decision trees are not foolproof, however, and may not work with some types of data. Some decision trees have problems handling continuous sets of data. Assigning values to groups in a "fuzzy" way can avoid this problem. Fuzzy tools are a subcategory of each category of tools. Fuzzy tools are most useful when a user is checking for multiple criteria and wants to vary the "closeness" of each criterion.

The basic philosophy of k-nearest neighbor is "do as your neighbors do." The behavior of a certain individual, for example, can be predicted by looking at the behavior of, for example, 10 individuals that are close to him in the data space. An average of the behavior of these 10 neighbors is calculated, and this average will be the prediction for the individual in question. The letter k in k-nearest neighbor stands for the number of neighbors investigated.

Neural networks resemble multiple regression in many ways except that, rather than using statistical theory as the basis of the technique, they imitate the information processing methods used by the human brain. Neural networks are, essentially, collections of connected nodes with inputs, outputs, and processing at each node. Between the visible input and output layers may be a number of hidden processing layers. The network has a training set of data for which the inputs produce a known set of outputs. Each case in the training set is compared with the known outcome. If it differs, a correction is calculated and applied to the processing in the nodes in the network. In other words, the network is capable of learning. The steps are repeated until a stopping condition, such as corrections being less than a certain amount, is reached. The resulting model does not have a clear interpretation and is usually applied without understanding the reasoning behind its results.

Some vendors recognize that different problems may be best-served by different approaches, so they combine these approaches and offer a suite of products. When applied properly, data mining can produce the return-on-investment from the data warehouse that managers have been waiting for. Some popular data mining tools are listed in Table 6.

Table 6. Data mining tools.

Product	Vendor	Phone
BrainMaker	California Scientific Software	800-284-8112
BusinessMiner	Business Objects	800-705-1515
Clementine	Integral Solutions	44 1256 55899 (UK)
Darwin	Thinking Machines	617-276-0400
Data Mining Suite	Information Discovery	310-937-3600
DataMind	DataMind Corp.	800-273-7102
DbProphet	Trajecta	800-250-2242
Decision Series	NeoVista	408-777-2929
//Discovery	Hyperparallel	415-284-7000
FuzzyTECH	Inform Software	630-268-7550
Intelligent Miner	IBM	800-426-2255
Knowledge Seeker	Angoss	416-593-1122
Mineset	Silicon Graphics	415-960-1980
SPSS CHAID	SPSS	800-543-2185

Data mining tools model the database to find relationships in the data. First, the business problems are identified. For example, a manager might want to find patterns to help retain good customers. Next, he/she identifies the types of models and the right data mining tools needed to answer specific questions. A regression model to forecast profitability might be built, as well as a classification model to categorize customers. The complexity of the problem will determine how difficult it is to extract meaningful relationships from the data. Complexity problems increase as the amount of data increases. Other contributors to complexity are the level of interaction among variables being examined and nonlinearity in the variables and parameters.

Data Visualization Techniques

Visualization techniques are very useful for discovering patterns in data sets. End users are presented with a display of the objects in their universe. They can mix and match these objects by pointing and clicking with the mouse to create the queries, reports, and graphs wanted. These may be used at the beginning of a data mining process to get a rough feel for the quality of the data set and where patterns are to be found. An elementary technique that can also be of great value is the scatter diagram. These diagrams can be used to identify interesting subsets of the data on which managers can focus the rest of the data mining process.

A two-dimensional visualization technique is displayed in Figure 9. In this example, a projection has been made along two dimensions: cost of service orders

and age of buildings. On average, service orders tend to be low cost in family housing with relatively new buildings.

A much better way to explore a data set is through an interactive three-dimensional environment. By viewing records as points in a multidimensional data space, analysts can see that records that are close to each other are alike, and records that are far from each other have little in common. Sometimes interesting clusters can be identified merely by visual inspection. In most cases, however, more advanced search programs are needed to uncover such clusters. Interesting predictions can also be visualized in this way. Sometimes it is possible to identify a visual cluster of potential high-cost projects using multidimensional space. In the Figure 9 example, age of buildings, construction type, and cost of service orders form an ideal three-dimensional space in which to do clustering analysis.

Figure 9. Cost of service orders and age of family housing buildings.

Modern visualization techniques can do much more than simple statistical graphs. The purpose of visualization is to transform data into information that is a critical component within the decisionmaking process. Data visualization can help managers quickly uncover and analyze trends and patterns in their

data. Managers then use this information to make more timely and informed business decisions. The interactive interface enables even the novice user to grasp and manage large database environments and drill down to the needed information. Data mining and data visualization work especially well together to solve business problems. Table 7 lists some popular data visualization tools available commercially.

Tools such as MineSet and the SAS system combine data mining with visualization. In such tools, visualizers work directly with data to explore relationships and trends using data mining algorithms. The tools present several dimensions of data simultaneously by using color, size, and animation. The visualizers support filtering, querying, rotation, zooming, and panning. Data mining methodology integrated with data visualization is a particularly useful feature that can help users integrate the steps of knowledge discovery.

Table 7. Data visualization tools.

Product	Vendor	Phone
Axum™	Mathsoft	800-548-5653
d.b.Express	Computer Concepts Corp.	800-619-0757
Data Desk	Data Description	800-607-1000
Discovery	Visible Decisions, Inc.	416-864-3900
Gific	LMI	800-374-4342
MineSet	Silicon Graphics Inc.	415-960-1980
NetMap	Alta Analytics	800-872-7144
PV-Wave	Visual Numerics, Inc.	800-447-7147
SAS	SAS Institute	919-677-8000
SPSS Diamond	SPSS	800-543-2185
Vis. Data Explorer	IBM	800-426-2255
WinViz	Information Technology Institute	65 778-7951 (Singapore)

6 Data Warehouse for Army Installations

Current Climate at Army Installations

Army installations use many DBMS and operational systems to conduct business. Development of applications for Army installations follows the traditional approach with limited scope and well-defined requirements, with a focus on separate functional areas. This approach creates the spider-web architecture between applications to integrate information. This architecture is very difficult and costly to maintain.

The installation commander must be able to reach across all the activities and functions on the installation to provide effective leadership and management. Existing systems designed to support specific functions such as finance, supply, or housing are not integrated with one another nor are they designed to give the local commander essential management information. Business practice improvement efforts often require supporting integrated business processes that cut across existing organizational lines. The data warehousing mechanism consolidates data from departmentally focused systems into a central repository that supports a cross-organizational viewpoint.

Besides the problems and difficulties associated with a spider web of extract programs, Army installations face some specific problems. The Army's use of COTS systems presents a new problem of how to manage Army corporate data independent of vendor modifications to their database schemas as a result of product upgrades. The Army's tendency towards privatization of many services poses a new information technology challenge of how to share information practically with external customers, service providers, and privatization partners if they use information management systems that differ from the Army's.

Data extraction and sharing from these third-party systems can be much simpler with a data warehouse architecture than with traditional spider-web architecture. Data warehousing enables data to be joined from different sources for analysis in new and innovative ways. Within the government sector, initiatives to improve the quality of products and services has called attention to disparities in data, and the difficulties in obtaining accurate and timely answers to questions. The data warehouse provides improved performance, better data quality,

and the ability to consolidate and summarize data from heterogeneous legacy systems. The data warehousing mechanism improves the quality of data by defining common data structures and formats, and enforcing consistent data domain values through data transformation.

The HQDA Data Warehouse being developed by SACC will support the HQDA with consolidated data on key business areas (e.g., Army units, personnel, logistics, budget, readiness, facilities, etc.). The warehouse is an integrated environment with a central repository of detailed and summarized historical data. Techniques such as data mining, historical analysis, and trend analysis are used to access and analyze this historical data. Army installations lack such a data warehouse with integrated and historical data. Although IFS-M attempts to integrate information into a common database, it uses program interfaces between applications, which creates a spider-web situation, and no historical data are available for analysis.

The Army developed systems to estimate condition codes for facilities (BUILDER, PAVER, etc.), but facilities inventory data are not integrated with these systems. Maintenance task databases such as Maintenance Resource Prediction Model (MRPM) task data and R.S. Means are independent of facilities condition codes. Data are available for facilities component cost estimation (Unit Price Book), component condition codes (BUILDER, PAVER, RAILER, etc), maintenance tasks for components (MRPM, R.S. Means), but these databases are not integrated to provide decision support information. With integrated standard data elements in an installation data warehouse, the management will have access to integrated data to support facilities management decisions.

BUILDER component and CACES (a cost estimation system) component structures do not follow the same format. MRPM follows CACES structure and has maintenance task data but is not integrated with BUILDER to show the relationship of condition codes and maintenance tasks. If all this data were available in a data warehouse, users could apply OLAP tools to see relationships between condition codes, maintenance tasks, and maintenance costs. This type of analysis will help management estimate maintenance costs, budget, and develop an annual work plan.

Figure 10 shows the current approach for the facilities M&R functional area. IFS-M or COTS systems are being used to create M&R operational data, which is independent of condition codes and maintenance tasks data. Facility condition codes are being generated in engineered management systems. Installations use M&R cost data from the Means or MCACES databases. Maintenance tasks are developed for building systems and components in the MRPM database. All

these databases are independent. Under this structure, it is difficult to find useful information for decision support.

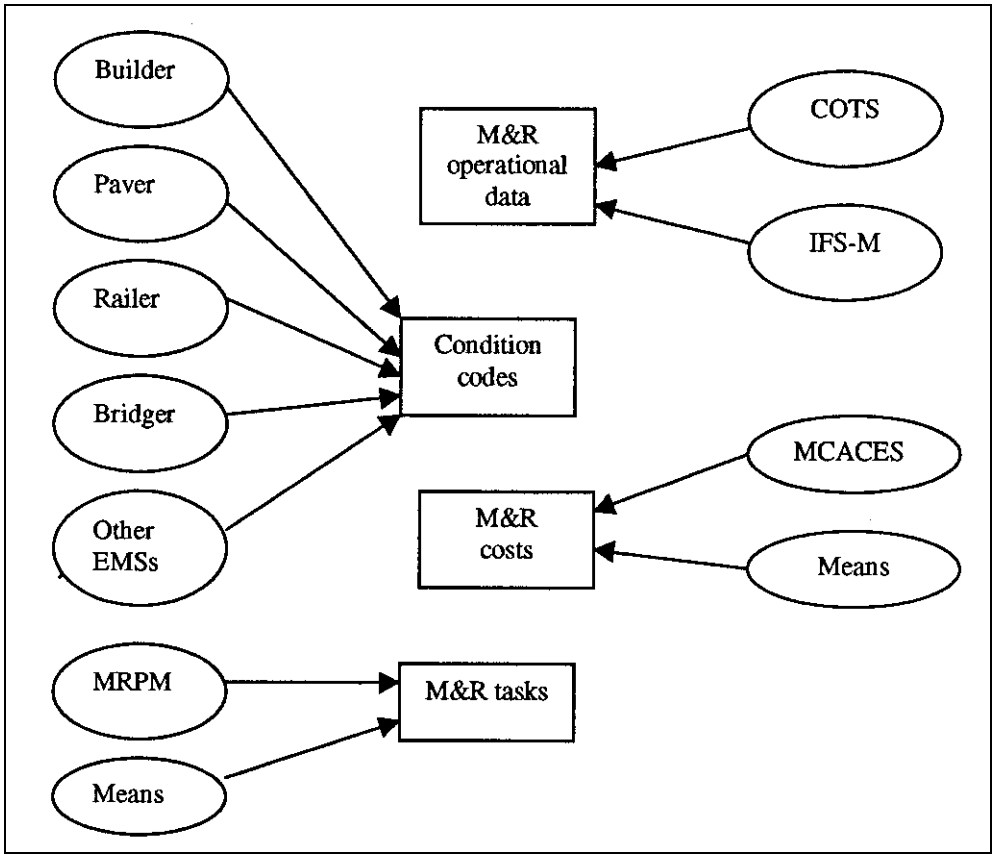


Figure 10. Facilities M&R data domain using current approach.

An Alternative Data Warehouse Approach

An alternative data warehouse approach is shown in Figure 11, which displays the warehouse's information flow architecture. Data for M&R operations, facilities condition codes, M&R tasks, and M&R costs are extracted, cleansed, integrated, and transported into the data warehouse. Additional time basis, summary, and derived data are added for analysis.

As a result of this data warehouse, management can access useful decision support information such as M&R cost effectiveness and evaluations of alternative M&R methods. This helps management choose optimum resource allocation methods for M&R actions.

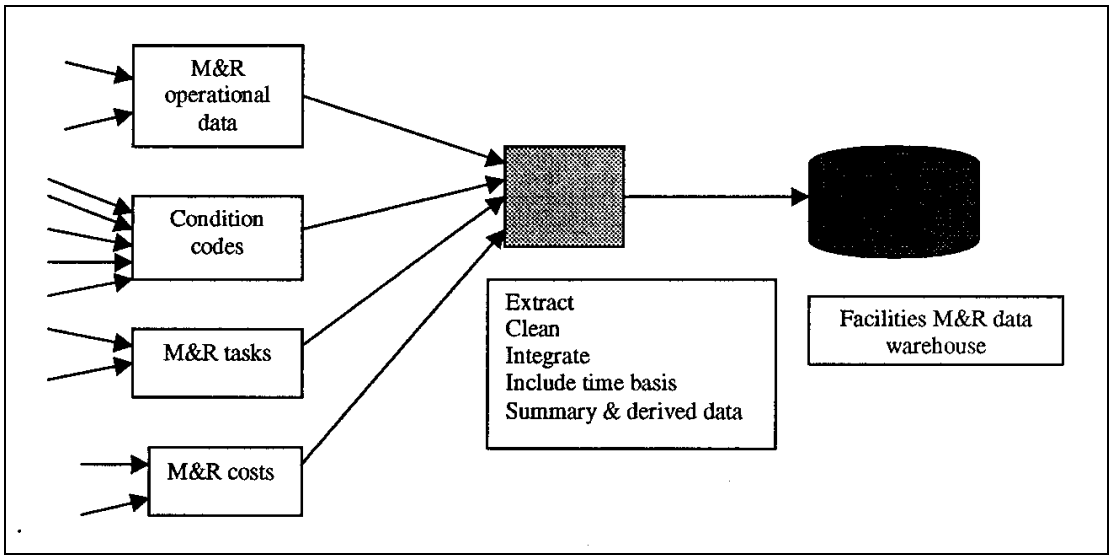


Figure 11. Facilities M&R data domain using the data warehouse approach.

Besides a lack of data integration in the current approach, a second major obstacle for decision support is that not enough historical data are stored in the applications. Operational systems were never designed to hold the historical data needed for decision support. Consider the following requests for information:

"How has the overall condition of facilities been different this year compared to the past 5 years?"

"Is there a trend in the M&R expenses during the past 10 years?"

"What are the lessons learned in the hazardous material management over the past 5 years?"

"What is the capacity of facilities/utilities/roads compared to requirements over a certain period?"

Going to existing operational systems to answer these questions is not an option because they lack integration and historical data.

To summarize the characteristics of the data warehouse:

1. The scope of the data warehouse eventually encompasses the whole enterprise.

2. The data warehouse contains the historical records of the business.

3. The source for all data in the warehouse is existing data from legacy systems.

4. The data in the warehouse is cleaned, transformed, and integrated using standard data elements.

The installation data warehouse can be a single source of information for the development of annual work plans, resource management plans, and long range plans. The data quality is improved with no inconsistencies and lag in time frame.

COTS systems will be like any other legacy system providing needed information to the data warehouse. COTS systems can be easily implemented as long as the required information is provided. The data will be cleaned, transformed, and integrated into standard data elements, and analysis and estimations will be more accurate.

Design and Development of an Installation Data Warehouse

Data warehousing is revolutionizing the way Federal Agencies access and use information. Government organizations are consolidating disparate databases that run on incompatible computer systems and forming centralized data repositories that enable quick information retrieval. Specialized tools are being used to manipulate data in those repositories to real patterns that can help executives make decisions and ultimately boost efficiency and cut costs.

Enterprise data modeling is the design technique that defines the contents of the warehouse to allow the entire scope of the installation business to be included in the data warehouse. The design of any data warehouse begins with the enterprise data model.

The focus of enterprise modeling is a complete and integrated view of all the data in the business. It aligns with the installation functional management structure rather than the data model of any application (legacy system). Application-level modeling provides a logical view of the data required by the application. It provides no significant support for integrating applications or for combining data from different sources. Supporting this combination of data from different sources requires the broader enterprise modeling. It outlines the logical and physical structure of the data warehouse with an integrated view.

To capture a highly consolidated view of an installation, primary functional areas such as facilities planning, design, programming, M&R, cost estimation, supply management, scheduling, and contract management are identified. Installations must outline clear objectives and anticipate expansion. Eventually,

other functional areas such as environmental compliance, pollution prevention, land management, and training management should be included in a data warehouse. The scope will cover the whole enterprise and all parts of the organization must be involved. Experts recommend starting with small, manageable applications, so a good approach is to tackle the problem piece by piece, rather than as one large effort. The Fort Eustis project, for example, began with developers extracting, cleansing, and transferring data from an IFS system into an operational data mart so that it is now an integrated system.

Once the initial subject areas are identified, define the contents of the different concepts in those subject areas in more detail. It helps to recognize the commonality between subject areas. For example, the relationship between M&R, facilities inventory, condition codes, resource management, and safety aspects needs to be recognized at a high level to ensure that data at lower levels in the model are correctly interrelated. Identification of commonality provides a link from the concepts to the generic ERM.

The ERM is a high level of modeling at the entity relationship level. It describes all data commonly used throughout the organization, and is developed to be enterprise-wide in its scope and generic to all of the application views. An ERM to cover the base operations functions of an installation can be considered an enterprise model for base operations of an installation. Since Army base operations are part of HQDA operations and DOD operations, this enterprise model needs to be consistent with the DOD enterprise model and HQDA data warehouse model. Standard data elements of DOD and Army need to be used in this effort to ensure the installation base operations ERM is consistent with other models.

Some entities will be used almost exclusively by one part of the organization. Such localized entities can be defined separately from the overall enterprise-wide aspects of the model. This compartmentalization is a practical approach to implementation, allowing the model to be defined in stages. For example, “exit” may be a localized entity for a legacy system available to safety professionals, and a “building category” entity may be shared by several users throughout the enterprise. The enterprise model supports the data warehouse design and provides overall business value to the end users. So, end-user input is required to develop the common and local data elements. Data definitions created as part of enterprise modeling form part of the metadata in the warehouse.

The scope of integration defines the boundaries of the model and determines what entities belong in the model. The scope of integration needs to be defined before the modeling process commences.

The first step in enterprise model development is to remove data used purely in the operational environment. Next, the key structures of the legacy systems are enhanced with an element of time. Derived data is also added where it is publicly used. Preventive maintenance projects by Facility Category Group (FCG) are derived data. An end user can see the changes in the real property inventory by year with an element of time added to the model.

Low-level modeling (the physical model) looks like a series of tables connected with arrows. The tables include keys and attributes for each entity. These data modeling methods are used to identify the attributes of data and the relationship between those attributes.

Another design activity is to perform stability analysis. Stability analysis is the act of grouping attributes of data together based on their propensity for change. Data that seldom change are grouped with like data, data that sometimes change are grouped with data that sometimes change, and data that frequently change are grouped with data that frequently change. The net result of this analysis is to create groups of data with similar characteristics.

Performance characteristics also need to be factored in while designing the physical database. Granularity and partitioning of the data, and creation of star joins to streamline data for DSS processing are some of the other design issues to be considered during the development of a data warehouse. There are many other physical design issues, most of which center around the efficiency of access to data.

While data architecture and modeling activities constitute a significant challenge in the design phase of the warehouse, the data transformation function often comprises the most costly and time-consuming part of the entire implementation. The process of data transfer and the steps in the process need to be well-defined during the data warehouse design phase. Source and target data requirements must be identified through data models. Mapping between source and target data must be created. This mapping is the first and most important requirement in data transfer strategy.

Data are transferred, for example, from the M&R portion of the IFS-M system or any other COTS or legacy system to Army standard data elements using the data model and star schema of the data warehouse. Data are transferred from other functional systems in a similar manner. Data transformations are done at field level and sometimes at table level. Metadata is created and made available in an easily accessible format to data warehouse administrators and end users. Most data warehousing construction tools will maintain a repository containing

this information. Data extraction and transformation tools typically contain operational metadata for administrators, while data query and presentation tools contain end-user metadata.

Quality data conversions are important to the data warehouse because the warehouse holds the information that is key to any decisionmaking process. Before beginning the conversion process, the data warehouse team completes the warehouse design, the physical data model for the warehouse, and generates the target schemas. Then a data conversion plan is developed. The conversion plan determines the best route by which to migrate source data to the data warehouse. The plan documents each source system's platform, access method, and language required for or extraction tool selected for data extraction. The data conversion plan will outline the appropriate strategy for gathering the data extractions to a common staging area to condition, clean, transform, and integrate the data. The strategy must take into account the volume of source data and available machine resources. Since a data warehouse must be built from source relational databases as well as nonrelational databases, the plan should cover a strategy for accessing data from different operational systems.

The first iteration of the data warehouse should be small enough to be built and large enough to be meaningful. The feedback loop between the warehouse developer and the end users at Army installations will constantly modify the warehouse data by adding other data to the warehouse. The time dimension is added to the data in the warehouse, because it must provide a historical view of the business. Probably the most obvious characteristic of historical data is its potential volume and the associated costs. Summary data is generally used over a longer time span than detailed data. At some point in time, data is purged from the warehouse. The issue of purging data is one of the important design issues of a data warehouse.

Hardware and Software Platforms

The platform selection is a matter of site preference as long as it meets the minimum requirements of the tools selected. Data warehouses can be built according to several different architectures. Early data warehouse implementations were built on large mainframe systems. In today's client/server and distributed environments, mainframe computing is no longer needed. Several data warehousing tools are available to support client/server technology, object-oriented database component (ODBC) technology, and dimensional modeling. For example, the Informatica® PowerMart suite supports source analysis, extraction, transformation design, mapping, multidimensional schema design,

warehouse design, repository management, and server management in an integrated tool set. PowerMart clients are compatible with both Microsoft® Windows® 95, and Windows® NT. Based on the tools and site requirements, servers and databases need to be chosen.

Client/server systems are installed in various forms: centralized, distributed, three-tier architecture, and multi-tier architecture (see Chapter 3, Data Warehouse Architecture From a Client/Server Perspective). Operating systems and database management systems with open, standardized interfaces create the conditions for heterogeneous linked systems. For data warehousing implementations, the data sources are fixed, and the data warehouse and server locations and platforms are flexible. A good design sequence is to first identify the data sources and types. Then choose the target data warehouse hardware and RDBMS platform. The number of users, type of query tools, and the volume of expected data in the data warehouse also influence the hardware selection. The conceptual data architecture, for data warehouse and metadata architecture, is the basis for hardware selection to support both the data management needs of the administrators and the data access needs of the end users.

Information Access and Decision Support

The primary purpose of the data warehouse is to provide consistent, understandable, decision support information to the end users. Multiple analysis procedures on consolidated and integrated information in a data warehouse use a goal-directed framework and thus serve the needs of management in the decisionmaking process. A wide variety of tools exist today to provide different analysis capabilities: query and reporting tools, OLAP tools, and data mining tools.

Most *query and reporting tools* focus strongly on the process of building the query or procedure for an individual user. These tools have evolved from the original need to put a more understandable face on the data access language. While the user interface of such tools has greatly improved over the years, especially with the advent of Windows-based front ends, these tools are usually more appropriate to users who are data literate. These tools are suitable only if the data requirements are simple and well understood in business terms.

OLAP tools provide end users with multidimensional analysis capability. Multidimensional analysis enables operations such as drill-down, roll-up, cross tabulations, ratios and trends, slice and dice, and data pivoting. A well-implemented multidimensional database allows an end user to quickly focus on the exact view

of the data required. The end user may want to determine how the maintenance costs of metal roofs compare to maintenance costs of asphalt shingle roofs. A manager might want to compare building maintenance costs with site maintenance costs. Another end user may want to focus on a specific problem such as condition of a training facility over years in order to detect trends.

Trend analysis and comparisons are not the only methods accommodated by OLAP. For the "slice and dice" approach, the analyst takes basic information grouped in one way and analyzes it; then the data is grouped another way and reanalyzed. In this way the manager can determine where the troublesome trends are. Many EIS are developed using OLAP tools.

Some knowledge hidden in the databases may be hard to find using query and reporting tools or even OLAP tools. For example, the objectives of an real property management and analysis tool (RMAT) project are to analyze the carrying capacity and to determine capital investment strategy. This type of information is not readily available in the databases. The capacity and strategy type information can be calculated, estimated, or predicted using knowledge discovery methods and analysis and prediction methods. The knowledge discovered from the databases can become the foundation for strategy development. *Data mining tools* such as statistical techniques, classification, clustering, decision trees, and neural networks support knowledge discovery by finding patterns in the data and inferring rules from them. These patterns and rules can be used to guide decisionmaking.

The potential payoffs from data mining are enormous if the right tools are chosen and used effectively. Intelligent data mining requires a mixture of tools and techniques. Some vendors combine approaches and offer a suite of products in recognizing that different problems may be best served by different approaches. Chapter 5 has more information on data mining.

7 Summary and Conclusions

A data warehouse is a consolidated database designed to facilitate data access from multiple data sources. Many of the information systems without a data warehouse are legacy systems that use and create data to produce local functional information with limited scope, not shared information. Reprogramming legacy systems for shared data access is expensive and risky. None of the alternatives to the data warehousing approach offers a complete solution to information sharing and decision support information access problems. Instead of an application with which users interface directly, a data warehouse is an integral part of an organization's underlying technical infrastructure.

The adoption of a data warehousing approach has helped many companies respond to an ever-shifting competitive environment. Most large companies have installed data warehouses, or are in the process of doing so. Government organizations, like those in the private sector, are either building data warehouses, considering building them, or involved in a transition from older technologies to client/server technology. The U.S. Army SACC, is developing a data warehouse to support the HQDA with consolidated data on Army units, personnel, logistics, facilities, readiness, and budget. Army installation data warehouses, when built, should be consistent with the HQDA data warehouse with standard data elements, naming conventions, formats, and definitions.

Army installations use many DBMS and operational systems with limited scope. These systems lack integrated and historical data. The current approach used by the Army to support installation management is to develop program interfaces between applications to integrate information. The data warehousing approach enables data to be joined from different sources and analyzed in new and innovative ways. The result is improved performance, better data quality, and the ability to consolidate and summarize data from heterogeneous legacy systems. It consists of historical data that can be used for trend analysis, pattern recognition, and prediction with data mining techniques. The installation data warehouse can be a single source of information for the development of annual work plans, resource management plans, and long range plans.

The enterprise data modeling design technique defines the contents of the warehouse to allow the entire scope of the installation business to be included in the

data warehouse. The focus of enterprise modeling is a complete and integrated view of all the data in the business. It aligns with installation functional management structure rather than the data model of any legacy system. Primary BASOPS functional areas to be integrated include facilities planning, design, programming, M&R, cost estimation, scheduling, environmental compliance, and supply, contract, land, and training management.

By recognizing the commonality between subject areas, the contents of the different concepts in those subject areas can be defined in more detail. Identification of commonality provides a link from the concepts to the generic ERM. The ERM describes all data commonly used throughout the organization. An ERM to cover the base operations functions of an installation can be considered as an enterprise model for base operations of an installation. The next step is to design the physical data model for the data warehouse and the target schemas.

Building a data warehouse requires data population from sources to target. Data models are used to identify source and target data requirements. Other details such as target database, data transfer and mapping strategy, and hardware and software to support data extraction, mapping, transfer, and loading are clarified and defined during this phase of the data warehouse development. Once the mapping between source and target has been created, data can be transferred between source and target based on the defined mapping. Data mapping takes care of data accuracy and format issues.

While enterprise data modeling is the primary effort in the design phase of the warehouse, data mapping and data loading are the primary efforts of the implementation phase of the data warehouse. Several commercial tools are available to automate parts of this implementation effort. All data shared throughout the organization and other consolidated information to support decisionmakers is extracted, transformed, and loaded into the data warehouse from all source legacy systems. A time element is added to the data in the data warehouse to access historical information and to perform analysis.

Metadata is created and made available in an easily accessible format to data warehouse administrators and end users. Metadata describes the meaning and structure of business data and of the corresponding application functions. Metadata is important to the data warehouse because it is through metadata that the data is registered, accessed, and controlled in the warehouse environment. Most data warehousing construction tools will maintain a repository containing this information. Data extraction and transformation tools typically contain operational metadata for administrators, while data query and presentation tools contain end-user metadata.

The first iteration of the data warehouse should be small enough to be built and large enough to be meaningful. The feedback loop between the warehouse developer and the end users at Army installations will constantly modify the warehouse data, adding other data to the warehouse. The time dimension is added to the data in the warehouse to provide a historical view of the business, but this historical data can become voluminous and thereby costly. The solution to this problem is to condense historical data from daily to weekly, monthly, or even yearly. At some point determined by management, data is purged from the warehouse. If the volumes of data are not carefully managed and condensed, the sheer volume of data that aggregates in the data warehouse prevents the goals of the warehouse from being achieved.

The key purpose for the data warehouse is to provide consistent, understandable, decision support information to end users. Multiple analysis procedures on consolidated and integrated information in a data warehouse use a goal-directed framework to serve the needs of management in the decisionmaking process. A wide variety of tools exist to provide different analysis capabilities: query and reporting tools, OLAP tools, and data mining tools.

Query and reporting tools are suitable for the analysis of simple and well understood data requirements. OLAP tools provide the multidimensional analysis capability to the end users. Multidimensional analysis allows end users to perform operations such as drill-down, roll-up, slice and dice, and cross tabulations. Data mining tools support knowledge discovery by finding patterns in the data and inferring rules from them. These patterns and rules can be used to guide decisionmaking. The potential payoffs from data mining are enormous if used effectively.

Data warehouses provide a framework for decision support processing and support executive decision making. Once the data warehouse has been built, the job of providing decision support information is infinitely easier than when no foundation of data exists on which to operate. The data in the warehouse is cleaned, transformed, and integrated using standard data elements making it a single accurate source of information for all decision support processing. Data quality is improved with no inconsistencies or lag in timeframe. Current client/server technology provides a less expensive and practical hardware platform to implement a data warehouse, and many commercial tools are available to support the design and development of a data warehouse. If designed and developed properly, an Army installation data warehouse has the potential to improve efficiencies and produce a positive return-on-investment.

References

Cited

Adriaans, Pieter, and Dolf Zantinge, *Data Mining* (Addison-Wesley, 1996).

Demarest, Marc, "Building the Data Mart," *DBMS Magazine*, July 1994.

Devlin, Barry, *Data Warehouse, From Architecture to Implementation* (Addison-Wesley, 1997).

Inmon, W.H., *Building the Data Warehouse* (John Wiley & Sons, Inc., 1996).

Installation Corporate Information Management (ICIM) Strategic Plan, draft charter, 1995.

Uncited

Appleton, Elaine, "Data Warehouse with an OLAP View," *Datamation*, April 1996.

Army Regulation [AR] 420-10, *Management of Installation Directorates of Public Works* (Headquarters Department of the Army [HQDA], Washington, DC, April 1997).

AR 415-15, *Army Military Construction Program Development and Execution* (HQDA, Washington, DC, August 1994).

Army Technical Manual 5-800-4, *Programming Cost Estimates for Military Construction* (Headquarters Department of the Army [HQDA], Washington, DC, May 1994).

Asbrand, Deboraw, "Is Datamining Ready for the Masses?" *Datamation*, November 1997.

Beitler, Stephen, and Ryan Leary, "Sears' EPIC Transformation: Converting from Mainframe Legacy Systems to On-Line Analytical Processing (OLAP)," *Journal of Data Warehousing*, April 1997.

Bohn**,** Kathy, "Converting Data for Warehouses," *DBMS Magazine*, June 1997.

Brooks, Peter, "March of the Data Marts," *DBMS Magazine*, March 1997.

Brooks, Peter, "Data Mining Today," *DBMS Magazine*, February 1997.

D. Appleton Company, Inc., "CIM Process Improvement Methodology for DOD Functional Managers," CIM training manual, January 1993.

Darling, Charles, and William Semich, "Wal-Mart's IT secret: Extreme Integration," *Datamation*, November 1996.

Darling, Charles, "How To Integrate Your Data Warehouse," *Datamation*, May 1996.

DBMS Magazine, Interview with Ralph Kimball, "Ralph Kimball Gets The Data Out," vol 7, no. 8, July 1994.

Department of the Army [DA] PAM 420-6, *Facilities Engineering, Resource Management*, Assistant Chief of Staff for Installation Management, May 1997.

Department of Defense [DOD] Directive 8000-1, *Defense Information Management (IM) Program*, ASD C3I, October 1992.

Devlin, Barry, "Managing Time in the Data Warehouse," *Journal of Data Warehousing*, Fall 1997.

DOD Directive 8320.1-M, *Data Administration Procedures*, ASD C3I, March 1994.

Haley, Barbara, "Implementing Successful Data Warehouses," *Journal of Data Warehousing*, Summer 1998.

Kay, Emily, "The Democratization of Datamining," *Datamation*, June 1998.

Kimball, Ralph, "A Dimensional Modeling Manifesto," *DBMS Magazine*, August 1997.

Kimball, Ralph, "Digging into Data Mining," *DBMS Magazine*, October 1997.

Lewison, Lisa, "Data Mining: Intelligent Technology Gets Down to Business," *PC AI*, December 1993.

Menninger, Dave, "Optimizing Your Data Warehouse for the Web," *Databased Advisor*, February 1997.

Neely, Edgar, R.D. Neathammer, and J.R. Stirn, *Maintenance Resource Prediction in the Facility Life-Cycle Process*, Technical Report P-91/10/ADA236424 (U.S. Army Construction Engineering Research Laboratory May 1991).

Noaman, Amin, and Ken Barker, "Distributed Data Warehouse Architectures," *Journal of Data Warehousing*, April 1997.

Park, Yong-Tae, "Strategic Uses of Data Warehouses: An organization's suitability for data warehousing," *Journal of Data Warehousing*, January 1997.

Sakaguchi, Toru, and Mark Frolick, "A Review of the Data Warehousing Literature," *Journal of Data Warehousing*, January 1997.

Schardt, James A., "The Enterprise Intersection Model: A unifying framework for data warehouse development," *Journal of Data Warehousing*, April 1997.

Schur, Stephen G., *The Database Factory: Active Database For Enterprise Computing* (John Wiley & Sons, Inc., 1994).

Tharpe, Maj. Leonard, "Data Warehousing in Headquarters Department of the Army (HQDA)," White paper, Strategic and Advanced Computing Center.

The, Lee, "OLAP Answers Tough Business Questions," *Datamation*, May 1995.

Watson, H.J., and B.J. Haley, "Data Warehousing: A framework and survey of current practices," *Journal of Data Warehousing*, University of Georgia, January 1997.

Whipple, Larry, "OLAPping at the Shores of Analysis," *Databased Advisor*, February 1997.

List of Acronyms

ACS(IM)	Assistant Chief of Staff (Installation Management)
CERL	Construction Engineering Research Laboratory
COTS	commercial off-the-shelf
DBMS	database management system
DPW	Directorate of Public Works
DSS	decision support system
EIS	executive information system
EPA	U.S. Environmental Protection Agency
ERD	entity relationship diagram
ERM	entity relationship model
FCG	Facility Category Group
HQDA	Headquarters, Department of the Army
ICAM	Integrated Computer-Aided Manufacturing
IDEF	ICAM Definition
I/O	input/output
ISD	Installation Support Division
M&R	maintenance and repair
OLAP	online analytical processing

OLTP	online transaction processing
RMAT	real property management tool
RDBMS	relational database management system
SACC	U.S. Army Strategic and Advanced Computing Center
SQL	structured query language

CERL Distribution

Chief of Engineers
ATTN: CEHEC-IM-LH (2)
ATTN: HECSA-Mailroom (2)
ATTN: CECC-R
ATTN: CEMP-IB (2)
ATTN: CERD-L
ATTN: CERD-M

Commander ERDC, Vicksburg, MS
ATTN: Coastal and Hydraulics Lab, Vicksburg
ATTN: Cold Regions Research, Hanover, NH
ATTN: Environmental Lab, Vicksburg
ATTN: Geotechnical Lab, Vicksburg
ATTN: Information Technology Lab, Vicksburg
ATTN: Structures Lab, Vicksburg
ATTN: Topographic Engineering Center, Alexandria, VA

Defense Tech Info Center 22304
ATTN: DTIC-O (2)

19
11/99